Describing the Statistical Conformation of Highly Flexible Proteins by Small-Angle X-ray Scattering

by

Jo Anna Wiersma Capp

Department of Biochemistry
Duke University

Date:_______________________
Approved:

____
Terrence G. Oas, Supervisor

____
David Richardson

____
Harold Erickson

____
Paul Modrich

____
Stefan Zauscher

Dissertation submitted in partial fulfillment of
the requirements for the degree of Doctor
of Philosophy in the Department of
Biochemistry in the Graduate School
of Duke University

2014

<u>ABSTRACT</u>

Describing the Statistical Conformation of Highly Flexible Proteins by Small-Angle X-ray Scattering

by

Jo Anna Wiersma Capp

Department of Biochemistry
Duke University

Date:_______________________
Approved:

____
Terrence G. Oas, Supervisor

____
David C. Richardson

____
Harold Erickson

____
Paul Modrich

____
Stefan Zauscher

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the Department of
Biochemistry in the Graduate School of
Duke University

2014

Abstract

Small-angle X-ray scattering (SAXS) is a biophysical technique that allows one to study the statistical conformation of a biopolymer in solution. The two-dimensional data obtained from SAXS is a low-resolution probe of the statistical conformation - it is a population weighted orientational average of all conformers within a conformational ensemble. Traditional biological SAXS experiments seek to describe an “average” structure of a protein, or enumerate a “minimal ensemble” of a protein at the atomic resolution scale. However, for highly flexible proteins, an average structure or minimal ensemble may be an insufficient representation of conformational space, and have more details than are justified by the information content of the data. This work describes a SAXS analysis of highly flexible proteins and presents a protocol for describing the statistical conformation based on minimally parameterized polymer physics models and judicious use of ensemble modeling. This protocol is applied to the structural characterization of *S. aureus* protein A (a crucial virulence factor) and Fibronectin III domains 1-2 (an important extracellular matrix protein in many higher eukaryotes). This work also discusses when it is appropriate to use a minimally parameterized polymer physics model and when using more-parameterized ensemble models is justified by the information content of the SAXS data.

Contents

List of Tables

List of Figures

Acknowledgements

I wish to thank all those who have supported me through my years of graduate school. This work would not have been possible without the assistance of members of the Oas, Richardson, and Erickson labs. Specifically, I wish to thank Yang Qi, Lindsay Deis, William Franch, and Andrew Hagarman for their assistance on the SpA-N project, Bradley Hintze, Gary Kapral, and Jeff Headd for their help with Python coding and Linux systems, and Riddhi Shah and Tomoo Ohashi for their assistance with the fibronectin project. Of course, I thank my advisors, Dave Richardson and Terry Oas - I wouldn’t be submitting this dissertation without their help and never-failing support. I would also like to thank my family and my husband, Chris Capp, for always being there to pick up the pieces when things fell apart.

1. Introduction to X-ray scattering

1.1 Introduction

Small-angle x-ray scattering (SAXS) is a fundamental biophysical method used in structural analysis of biological macromolecules and other biopolymers. The scattering of x-rays at small angles contains important information about the in-homogeneities in electron density of biological macromolecules with characteristic length scales between 1 – 100 nm (10 – 1000 Å) (Roe, 2000). Global structural parameters (volume, and radius of gyration) of biological macromolecules can be determined from SAXS analysis. More importantly, SAXS is sensitive to the internal structure of disordered, partially ordered, and highly flexible systems. Thus, it is an ideal technique for studying the shape and conformations of biological macromolecules in solution (Receveur-Brechot and Durand, 2012). This chapter will provide an introduction to SAXS and the fundamental principles of x-ray scattering will be derived and reviewed. In Chapter 2 modern methods of modeling biological macromolecules using information determined from SAXS experiments will be reviewed. The limitations of SAXS in molecular modeling will also be discussed, with a particular emphasis on determining the information content of SAXS data. Chapters 3, 4, and 5 describe the results of SAXS analysis of systems of highly flexible proteins. Finally, Chapter 6 provides a discussion of the future of SAXS-based structural modeling.

1.2 General properties of X-rays

X-rays are electromagnetic waves of wavelengths from 10^{-3} to 10 nm in wavelength (10^{-2} to 10^{2} Å). Normally, the wavelengths used to study biological macromolecules by scattering or diffraction are 0.05 – 0.25 nm (0.5 – 2.5 Å) (Koch et al., 2003). Although x-rays are electromagnetic radiation, with both an electric field and a magnetic field, only the electric field will be considered because this is the component of electromagnetic radiation that gives rise to scattering and diffraction. Thus, for the purpose of this discussion, x-rays are defined as an electric field oscillating in a single plane, where the plane is perpendicular to the direction of propagation (Figure 1).

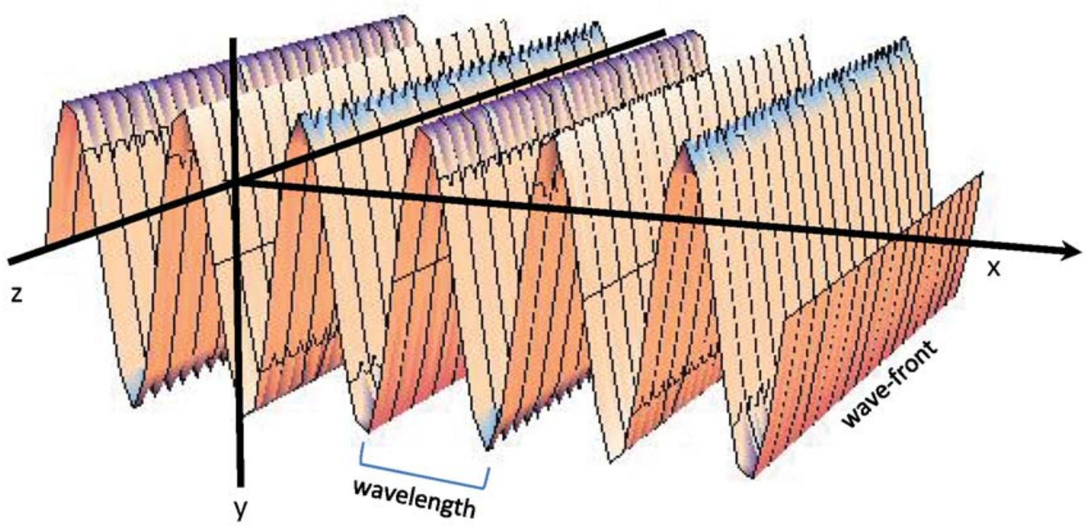

Figure 1: A planar wave. The wave-front, oscillating in the plane zy, is perpendicular to the direction of propagation (x).

When x-rays interact with matter the energy and amplitude of the x-ray are dissipated by the ejection of an orbital electron or by scattering. This dissipation results in several physical effects that can be detected by experimental methods (Roe, 2000):

1. Absorption of x-ray energy by the matter and ejection of an orbital electron
2. Emission of an x-ray of a characteristic wavelength (fluorescent x-ray)
3. Heat generation
4. Scattering of the x-rays

It is important to note that absorption and fluorescence result in a change in energy of the emitted x-ray (change in wavelength). These effects are due to the physical realities of the experiment – scattering is not completely elastic. However, for the purposes of this manuscript, scattering is considered to be elastic.

1.3 X-ray scattering – definition and description

X-ray scattering occurs when the incoming x-rays cause vibrations of the shell electrons in the atoms they encounter. This acceleration of shell electrons results in the emission of secondary radiation. Since the acceleration of shell electrons is stimulated by the oscillating electric field of the incident radiation, the scattered x-rays have the same wavelength as the incident radiation (Drenth and Mesters, 2007). In other words, the electrons oscillate at the same frequency as that of the incident radiation.

Before delving further into the description of x-ray scattering, there are important assumptions and simplifications about the scattering of x-rays that must be stated (Roe, 2000):

1. The scattered x-rays are coherent.
2. The scattered x-rays are propagated in three-dimensions over the whole of space. The shape of this propagation is dependent on the size and shape of the electron orbitals in the interacting matter. For now, the propagation of scattered x-rays is considered to be spherically symmetric.
3. Scattered x-rays do not interact with more than one scattering site. In other words, the scattered x-rays interfere with other scattered x-rays, but the observable effects of secondary interactions between scattered x-rays and the interacting matter are disregarded.
4. The incident x-rays are parallel, as are all the x-rays arriving at a specific observation point, given a large distance between the interacting matter and the observation point.

1.4 A diagrammatic description of scattering

When an x-ray beam interacts with a single particle of constant electron density, most of the incident photons will pass through the particle without being scattered. However, scattering of a small fraction of photons will occur in three dimensions

(Glatter and Kratky, 1982). To simplify the representation of the following diagrams, both the incident x-rays and the scattered x-rays are represented in two dimensions as wave-fronts. Depicted below (Figure 2) is a diagram of two-dimensional x-ray scattering. The incident radiation is depicted as a series of parallel plane waves (black), with the peaks of the waves represented by solid lines and the troughs represented by dashed lines. The scattered x-rays emanating from the interacting matter (blue box) are represented by the circular blue lines.

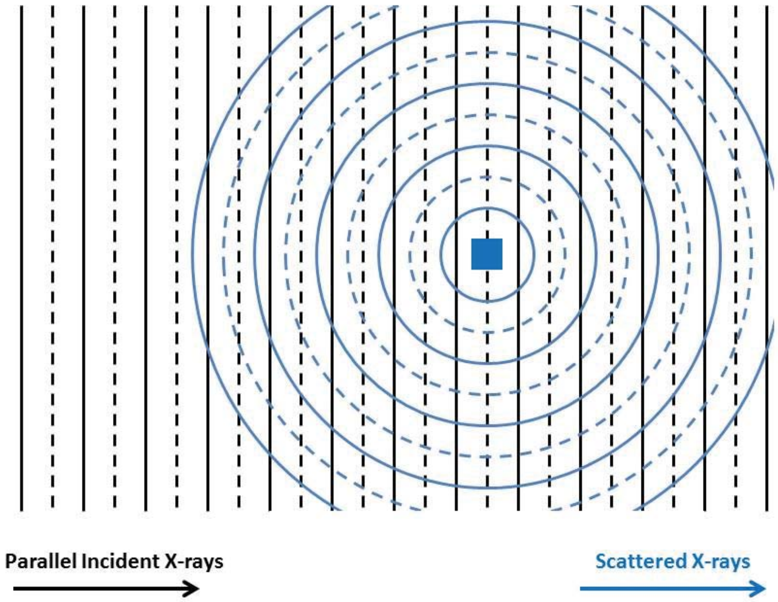

Figure 2: Two-dimensional diagram of scattering from a single object. Black: the incident x-rays, represented as a parallel wave-front. Blue: scattered x-rays as a spherically symmetric wave-front emanating from the center of the scattering material (blue box).

Waves can also be represented as vectors (Figure 3). In this case the planar incident x-ray (black), the transmitted x-ray (black), and the scattered x-ray (blue) are represented by vectors. The direction of the scattered x-ray is 2θ with respect to the incident x-ray. This results from the definition of the scattering vector (red). The scattering vector is defined as the difference between the incident x-ray and the scattered x-ray (Glatter and Kratky, 1982). Thus, to determine the magnitude of the scattering vector, the angle between the incident and scattered x-rays is bisected, such that two right triangles are formed. Then the magnitude of the scattering vector is simply $2\sin\theta$. This convention of defining the scattering angle as 2θ persists throughout the x-ray scattering literature.

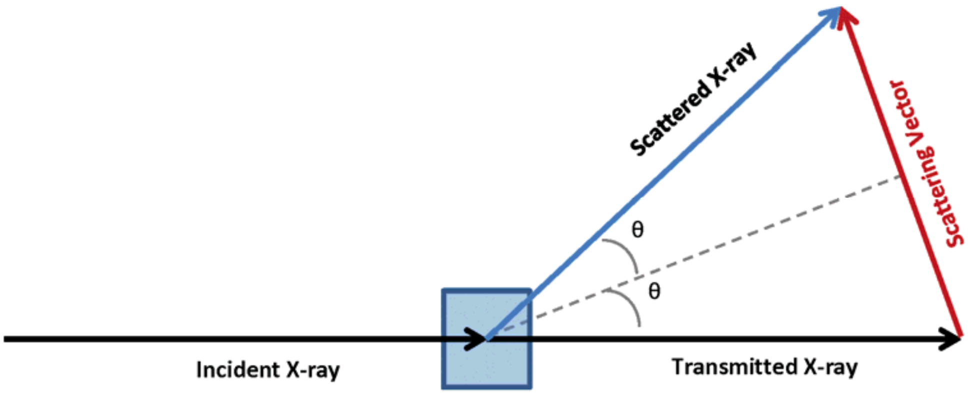

Figure 3: A vector depiction of scattering. The magnitude and direction of the scattered x-ray is given by the scattering vector.

It is useful to employ both the wave-front diagrams and vector depictions of scattering at this early stage. Later, we will exclusively use the vector representation in scattering diagrams for its simplicity.

1.5 Interference

Thus far, only scattering resulting from interactions of x-rays with a single point has been considered. However, in reality, we observe the scattering from electrons in many atoms within many molecules. The experimentally observed scattering from a molecule is the result of the scattering of x-rays from individual electrons and the interference among the waves scattered by these primary events (Roe, 2000). Figure 4 shows the two dimensional wave-front diagrams of scattering from a system of two identical atoms. The scattered spherical wave-front from the blue atom is in blue, and from the red atom is in red. Again, solid lines represent the peaks of the wave-fronts and dashed lines represent the troughs. Since the two atoms are identical, they will scatter the incident x-rays identically. In a system of two atoms, the scattered wave-fronts generated by the two atoms overlap – they interfere with each other. Vector **B** shows constructive interference: the peaks of the two scattered wave-fronts overlap. The amplitude of the resultant x-ray along **B** will be twice that from a single wave-front (Figure 4B). Likewise, along vector **C**, the two waves destructively interfere: the peak of one wave-front coincides with the trough of the other wave-front. In this case, the amplitude of the resultant x-ray along **C** will be zero (Figure 4C).

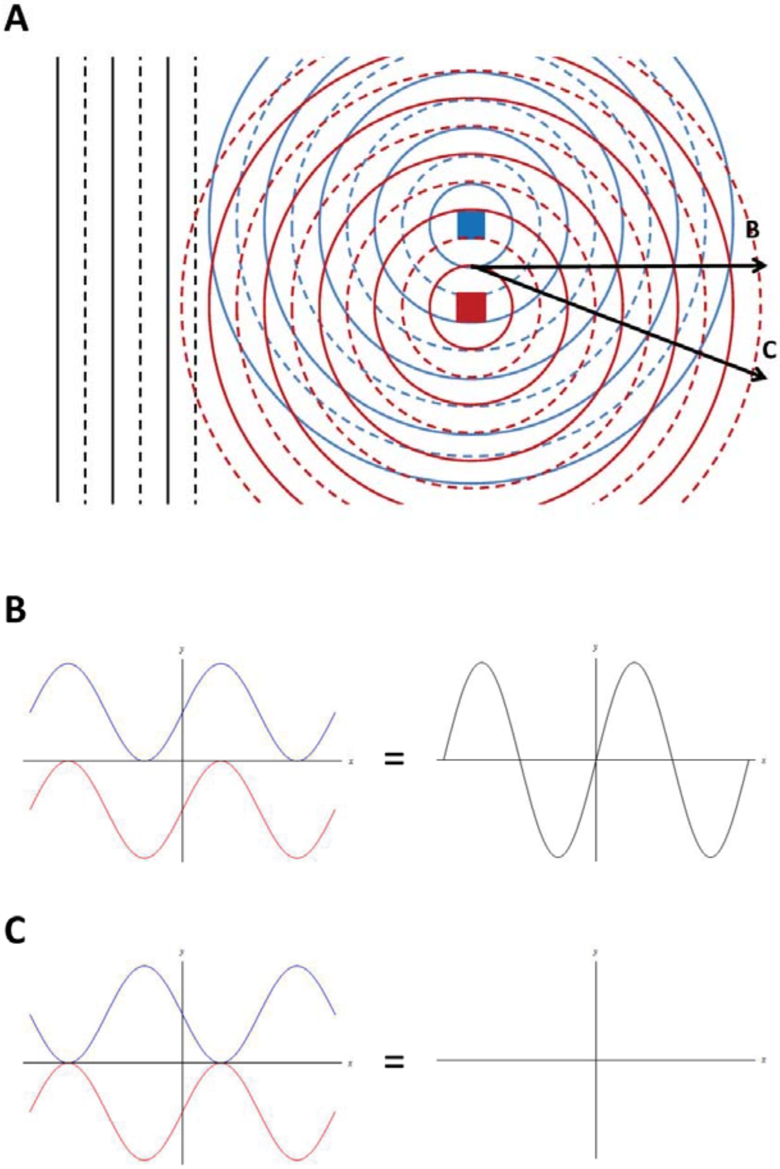

Figure 4: Scattering from a system of two identical atoms. A) Two dimensional wave-front diagram of scattering from two atoms. The two wave-fronts constructively

interfere along direction B, and destructively interfere along direction C. B) A wave diagram of the constructive interference along vector B. The two waves (blue, red) sum to give a resultant wave (black) that is twice the amplitude of the original waves. C) A wave diagram of destructive interference along vector C. The two waves (blue, red) sum to give a resultant wave with zero amplitude.

Figure 4B and C are examples of scattered x-rays from two points that are added together. In the case of 4B, the resulting wave is twice the amplitude of the individual waves. This is because the crest and trough of each wave coincide. The opposite is true in the case of 4C, the crest of each wave corresponds to the trough of the other wave. In 4C, the red wave lags behind the blue wave by ½ a wavelength. In angular terms, it lags behind the blue wave by π radians. In other words, there is a *phase difference* between the two waves (Drenth and Mesters, 2007). This phase difference has been expressed in terms of the angular difference between the associated circular motions (in units of $^{2\pi}/_{\lambda}$) (Roe, 2000). Vector C is not unique; the scattered waves in all directions interfere with each other to give a resultant wave that is the sum of the amplitudes and phases of the component waves. It is the difference in phases and amplitudes in all the scattered X-rays that allows shape specific information to be obtained from scattering and diffraction data.

1.6 Mathematical description of scattering

Until now, scattering has been discussed semi-graphically and qualitatively. This method, while it sets the basis for an intuitive understanding of scattering, has serious limitations when applied to the actual problem of data analysis. The following mathematical treatment of scattering from many atoms will reinforce the intuitive understanding of scattering developed in previous sections and will permit a deeper understanding of the relationships between the variables involved in scattering. First we will discuss how the amplitude and phase differences of scattered waves are calculated and then end with a discussion of the scattering form factor.

1.6.1 Calculation of amplitude and phase differences of scattered waves

The amplitude of a wave of frequency f and wavelength λ traveling in the x direction can be expressed as

$$A(t,x) = |A_0| e^{i2\pi(ft - \frac{x}{\lambda})} \qquad 1.1$$

$|A_0|$ is the absolute value of A(t,x). The term $\frac{2\pi x}{\lambda}$ gives the change in phase of a wave as it travels a distance x to an observation point. This general definition of a wave can be adapted to describe the scattering of an x-ray in relation to the incident x-ray. Since x-ray scattering is elastic and considered to be instantaneous, f and t may be disregarded (Roe, 2000). The amplitude of a scattered wave can be expressed as

$$A(x) = |A_0| b e^{-i\frac{2\pi x}{\lambda}} \qquad 1.2$$

In this case, $|A_0|$ is the absolute value of the incident x-ray, and b is the scattering efficiency. The magnitude of b depends on the composition of the particle. For example, an atom with a larger number of electrons will have a scattering efficiency that is larger than an atom with a smaller number of electrons (Roe, 2000). For now, it is assumed that the scattering particles have an identical scattering efficiency. When two scattering sites are present, the scattered x-rays interfere, and the amplitude of the resultant x-ray in the x direction will be the sum of the two scattered x-rays:

$$A_T(x_{12}) = A(x_1) + A(x_2)$$

$$A_T(x_{12}) = |A_0|be^{-i\frac{2\pi x_1}{\lambda}} + |A_0|be^{-i\frac{2\pi x_2}{\lambda}} \tag{1.3}$$

Since the incident x-ray and the scattering efficiency of the two scattering sites are identical, the only difference between the two x-rays is the path length, x_1 and x_2. The incident and scattered x-ray travels a distance that is dependent on the position of the scattering sites relative to each other. This difference in path length, Δx, is what results in a phase difference between the two scattered x-rays, $\frac{2\pi\Delta x}{\lambda}$. Determining the path length difference (and phase difference) lies at the heart of interpreting scattering data from complex macromolecules.

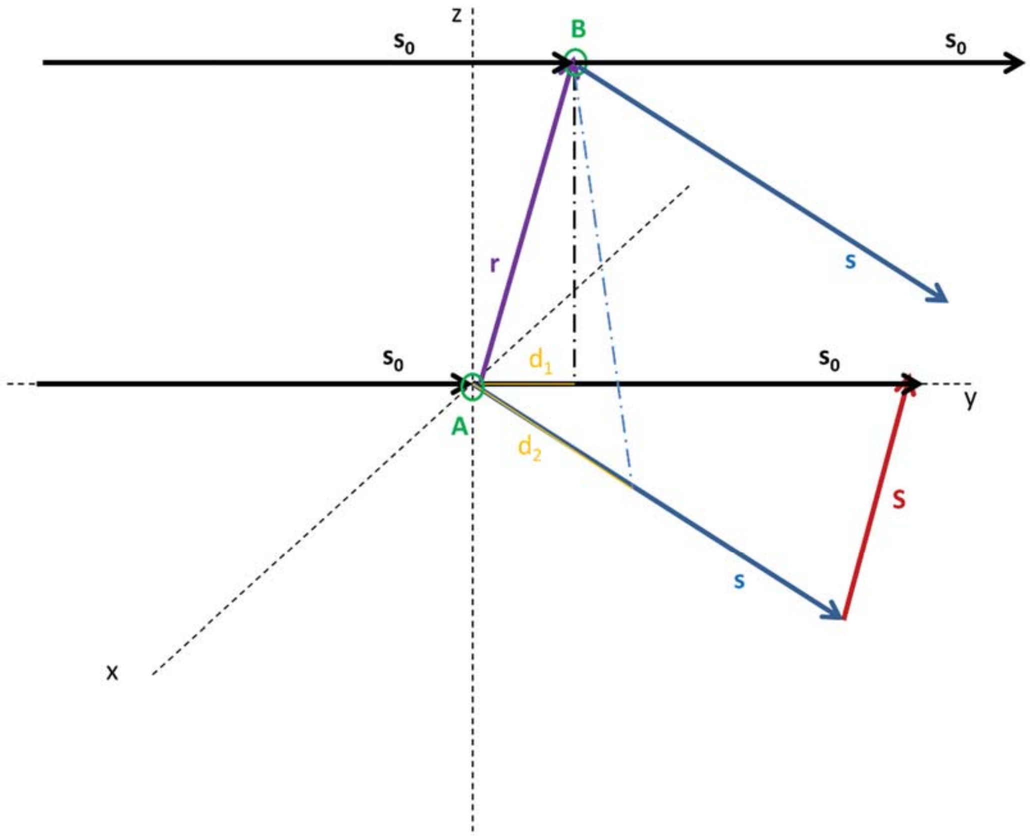

Figure 5: A three-dimensional vector diagram for scattering from two sites.

A vector diagram will be used to derive an expression for the path length difference and phase difference between two scattering sites (Richardson, 2014 – personal communication). Figure 5 shows the scattering between two points, A and B, a distance **r** apart. A is arbitrarily located at the origin. The incident x-ray, traveling in the direction specified by unit vector $\mathbf{s}_0$ (black), is scattered by points A and B (green circles) in the direction specified by unit vector **s** (blue). Vectors **s₀, s,** and **S** define a plane. The path length difference, Δx, is given by

$$\Delta x = d_2 - d_1 \tag{1.4}$$

d_1 is the projection of **r** on $\mathbf{s_0}$, and $d_2$is the projection of **r** on **s**. So,

$$d_1 = \boldsymbol{r} \cdot \boldsymbol{s_0} \tag{1.5}$$

$$d_2 = \boldsymbol{r} \cdot \boldsymbol{s} \tag{1.6}$$

$$\Delta x = \boldsymbol{r} \cdot \boldsymbol{s} - \boldsymbol{r} \cdot \boldsymbol{s_0}$$

$$\Delta x = \boldsymbol{r} \cdot (\boldsymbol{s} - \boldsymbol{s_0}) \tag{1.7}$$

Recall that the scattering vector, **S**, is defined as (from section 1.4)

$$\boldsymbol{S} = \boldsymbol{s} - \boldsymbol{s_0}$$

If this is substituted into equation 1.7, the path length difference between the two scattering sites becomes

$$\Delta x = \boldsymbol{r} \cdot \boldsymbol{S} \tag{1.8}$$

The amplitude of the resultant x-ray in direction x is therefore given by

$$A(\boldsymbol{S}) = |\boldsymbol{A_0}| b e^{-\frac{i2\pi \boldsymbol{r}\cdot\boldsymbol{S}}{\lambda}} \tag{1.9}$$

b is the scattering efficiency of the scattering site. So far, **S** has been defined as the absolute path length difference between points A and B. This is because **s** and $\mathbf{s_0}$ are unit vectors. To describe the relative path length difference between points A and B, $1/\lambda$ is applied to the unit vectors. The relative phase difference between the two scattering sites has been defined above as $\frac{2\pi\Delta x}{\lambda}$. Applying $1/_{\lambda}$to the unit vectors and substituting equation 1.8 in for Δx, (Roe, 2000)

$$\frac{2\pi\Delta x}{\lambda} = 2\pi\boldsymbol{r}\cdot\boldsymbol{S} \qquad 1.10$$

Equation 1.10 relates the path difference to the projection of r onto the scattering vector **S**. Now a new scattering vector, **q**, is defined**.** It converts the relative path length difference to the relative phase difference:

$$\boldsymbol{q} = 2\pi\boldsymbol{S} \qquad 1.11$$

equation 1.10 becomes

$$\frac{2\pi\Delta x}{\lambda} = \boldsymbol{r}\cdot\boldsymbol{q} \qquad 1.12$$

and equation 1.9 becomes

$$A(\boldsymbol{q}) = \ |\boldsymbol{A_0}|be^{-i\boldsymbol{r}\cdot\boldsymbol{q}} \qquad 1.13$$

Recall from section 1.4 that the magnitude of scattering vector **S** is 2sinθ. Therefore, the magnitude of **q** is (Guinier and Fournet, 1955):

$$|\boldsymbol{q}| = \frac{2\pi\ 2\sin\theta}{\lambda}$$

$$|\boldsymbol{q}| = \frac{4\pi\sin\theta}{\lambda} \qquad 1.14$$

Equation 1.14 defines **q** in reference to the scattering angle, 2θ.

A general expression for the summation of the phase difference between two scattering sites at a particular scattering angle, represented by **q**, has been derived (equation 1.13). This equation can be extended to a large number of scattering sites, N, and their phase difference relative to an arbitrary origin by

$$A(\boldsymbol{q}) = |\boldsymbol{A_0}| b \sum_{i=1}^{N} e^{-i\boldsymbol{r}_i \cdot \boldsymbol{q}} \qquad 1.15$$

for all i scattering sites in the particle (Roe, 2000). Again, b is the scattering efficiency of each scattering site, assuming all scattering sites are identical.

So far, the scattering sites have been treated as dimensionless points. However, in reality, whether we are considering atoms, or later, homogenous shapes, each scattering site has a definitive mass and electron density. In following sections this component of mass and electron density will be added to the amplitude equation.

1.6.2 The relationship between real and reciprocal space

An equation defining the amplitude of a collection of dimensionless points has been derived. However, the scattering volume of an object must also be considered. In equation 1.14 the combined amplitude of all point scatterers relative to an arbitrary origin is derived. Since interactions of x-rays with electrons is what gives rise to scattering, now we will discuss scattering in terms of the number of electrons within a particular volume – the electron density (Guinier and Fournet, 1955). If it is assumed that the number of scattering points is large and they are continuously dispersed within a particular volume, the sum in equation 1.15 may be replaced with an integral

$$A(\boldsymbol{q}) = A_0 \int_V \rho(\boldsymbol{r}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\, dr \qquad 1.16$$

Where $\rho(\boldsymbol{r})$ represents a number of electrons within a volume element: $dr = dx\ dy\ dz$. The units of $\rho(\boldsymbol{r})$ are therefore Å^{-3}, while dr has units of Å^{3}. This is also called the scattering length density distribution (Roe, 2000). V in the integration denotes that the integral is to be performed over a finite volume - the scattering volume. If we assume that the particle is made up of uniformly distributed electrons, the amplitude can be generally written as

$$A(\boldsymbol{q}) = \int_V \rho(\boldsymbol{r}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\, dr \qquad 1.17$$

where $A(\boldsymbol{q})$ is now understood to be the normalized scattering amplitude - the ratio of $A(\boldsymbol{q}) / A_0$. In the above equation (equation 1.17) A(**q**) is proportional to the three-dimensional Fourier transform of $\rho(\boldsymbol{r})$[1]. Thus, a Fourier transform of the volume of a scattering object will yield the amplitude of the scattering as a function of **q**:

$$\rho(\boldsymbol{r}) = \int_0^\infty A(\boldsymbol{q})\, e^{-i\boldsymbol{q}\cdot\boldsymbol{r}} dq \qquad 1.18$$

One important caveat in this equation is that the Fourier transform is only taken over a finite scattering angle, **q** (since experimentally q is finite), so in reality equation 1.18 is only an approximation of $\rho(\boldsymbol{r})$ that is dependent on the resolution of the experiment (Svergun, 1992).

---

[1] For an excellent explanation of Fourier transforms as applied to scattering please see Appendix B of "Methods of x-ray and neutron scattering in polymer science" (Roe, 2000), Oxford University Press.

Equation 1.18 allows us to define the relationship between real and reciprocal space in x-ray scattering. It has already been demonstrated that the magnitude of scattering vector,$|\boldsymbol{q}|$, has units of Å^{-1} while $|\boldsymbol{r}|$ has units of Å. Thus, these two vectors are reciprocally related: **r** is called the real-space vector, while **q** is the reciprocal-space vector. In other words, for a complex particle with a volume specified by $\rho(\boldsymbol{r})$, which is the set of all vectors **r** within the particle volume V, there will exist a set of vectors **q** in reciprocal space that fully accounts for all elements in **r**. One element in vector **r** in real-space will map to one element of vector **q** in reciprocal space. So, x-ray scattering and x-ray diffraction analysis can be described as a method to determine the pattern of scattering sites in real space that corresponds to the direction and amplitudes of the scattered x-rays observed in the reciprocal space experiment.

It is clear that since $\rho(\boldsymbol{r})$ and A(**q**) are related by a Fourier transform, they are interchangeable equivalent expressions (Figure 6) (Glatter and Kratky, 1982).

$$\begin{tikzcd}
\rho(\boldsymbol{r}) \arrow[r, shift left, "\text{Fourier transform}"] & A(\boldsymbol{q}) \arrow[l, shift left, "\text{inverse Fourier transform}"]
\end{tikzcd}$$

Figure 6: Relationship between real-space and reciprocal-space

1.6.3 The intensity calculation

So far, the amplitude of a particle with a specific volume has been determined. However, because of the nature of x-rays, the amplitude of the scattered waves cannot be experimentally observed. Instead, what is observed during the scattering experiment is the *intensity* of the scattered waves. The intensity of the scattered x-ray as a function of scattering angle is defined as (Glatter and Kratky, 1982)

$$I(q) = A(\boldsymbol{q}) \cdot A^{*}(\boldsymbol{q}) \tag{1.19}$$

$$I(q) = \left[\int_V \rho(\boldsymbol{r}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\, dr\right]\left[\int_V \rho(\boldsymbol{r}) e^{i\boldsymbol{q}\cdot\boldsymbol{r}}\, dr\right] \tag{1.20}$$

The intensity is the absolute square of the Fourier transform of the scattering volume density. Therefore, there is no way to directly convert I(q) directly to $\rho(\boldsymbol{r})$ (Figure 7). When $A(\boldsymbol{q})$ is multiplied by its complex conjugate to yield I(q) some of the information contained in $\rho(\boldsymbol{r})$ is lost. This is because the scattering amplitude is a complex quantity; when it is converted to intensity, information about the phase angle between each scattering site and the origin is completely lost. It is necessary to have a starting model for the scattering length density distribution (Roe, 2000). Most modern methods of scattering data interpretation rely on a good starting model(s), compare the scattering data to this model, then optimize the fit of the model to the data. These methods will be reviewed in detail in Chapter 2.

$$\begin{tikzcd}
\rho(\boldsymbol{r}) \arrow[r, shift left, "\text{Fourier transform}"] & A(\boldsymbol{q}) \arrow[l, shift left, "\text{inverse Fourier transform}"] \arrow[d, "\text{squaring}"'] \\
 & \mathrm{I}(\boldsymbol{q})
\end{tikzcd}$$

Figure 7: Relationship between $\boldsymbol{\rho(r)}$, A(q) and I(q)

1.6.4 Orientational averaging in small-angle scattering

So far, the scattering from a single oriented particle has been considered. Additionally, all the concepts presented above are common to x-ray scattering and x-ray diffraction. Now, the discussion narrows to focus solely on x-ray scattering of particles in solution.

Equation 1.17 is a general equation that defines the scattering amplitude as a function of the scattering length density distribution. The problem with this equation is that the orientation of vector **r** has to be specified with respect to **q** (Figure 5). However, the orientation of molecules in solution is not fixed. Molecules are able to freely rotate. The experimentally observed intensity is a super-position of all possible orientations of the molecule (Putnam et al., 2007). There is an important assumption in interpretation of x-ray scattering data: the solution must be isotropic. Therefore, **r** is integrated in three dimensions. An expression for the orientationally averaged scattering amplitude is derived below. If **r** is expressed in spherical polar coordinates in terms of r, Θ, and Φ, and (Roe, 2000)

$$dr = r^2 \sin\Theta \; dr \, d\Theta \, d\Phi \tag{1.21}$$[2]

then,

$$A(\boldsymbol{q}) = \int_{\Phi=0}^{2\pi} \int_{\Theta=0}^{\pi} \int_{r=0}^{\infty} \rho(\boldsymbol{r}) \, e^{-i\boldsymbol{q}\cdot\boldsymbol{r}} r^2 \sin\Theta \, dr \, d\Theta \, d\Phi \tag{1.22}$$

Since $\rho(\boldsymbol{r})$ is the real-space scattering length density distribution, A(**q**) is a real function of length **q**. If we choose the polar axis to coincide with the direction of **q**,

$$\boldsymbol{q} \cdot \boldsymbol{r} = qr \cos\Theta \tag{1.23}$$

Equation 1.23 changes $\boldsymbol{q} \cdot \boldsymbol{r}$ to a scalar quantity. If we assume that the isotropic averaging of the particle is centrosymmetric, $\rho(r) = \langle \rho(\boldsymbol{r}) \rangle$, then the average amplitude, A(q) = $\langle A(\boldsymbol{q}) \rangle$ also becomes a scalar quantity (Guinier and Fournet, 1955):

$$\langle A(\boldsymbol{q}) \rangle = b_e \int_{\Phi=0}^{2\pi} \int_{\Theta=0}^{\pi} \int_{r=0}^{\infty} \langle \rho(\boldsymbol{r}) \rangle \, e^{-iqr\cos\theta} r^2 \sin\Theta \, dr \, d\Theta \, d\Phi$$

$$A(q) = b_e \int_{\Phi=0}^{2\pi} \int_{\Theta=0}^{\pi} \int_{r=0}^{\infty} \rho(r) \, e^{-iqr\cos\theta} r^2 \sin\Theta \, dr \, d\Theta \, d\Phi \tag{1.24}$$

---

[2] Refer to Appendix A for this derivation.

If a u-substitution is done such that u = cos Θ, equation 1.22 becomes

$$A(q) = b_e \int_{\Phi=0}^{2\pi} \int_{u=-1}^{1} \int_{r=0}^{\infty} \rho(r)\, e^{-iqru} r^2 \sin\Theta \; dr\; du\; d\Phi \qquad 1.25$$

$$A(q) = 2\pi b_e \int_{r=0}^{\infty} \rho(r) r^2 \frac{e^{iqr} - e^{-iqr}}{iqr} dr \qquad 1.26$$

Recall that

$$\sin x = \frac{e^{ix} - e^{-ix}}{2i} \qquad 1.27$$

Substituting this expression into the equation,

$$A(q) = 4\pi b_e \int_{r=0}^{\infty} \rho(r) r^2 \frac{\sin qr}{qr} dr \qquad 1.28$$[3]

The average amplitude of the scattered x-rays as a function of scattering angle is determined by the distribution of scattering centers as a function of their distance from the center of the particle. It is this orientational averaging that makes scattering distinct from crystallography.

1.6.5 Scattering from polyatomic molecules

Scattering from a collection of dimensionless points and scattering from an object with uniform electron density have been discussed. Now, the equation for the scattering amplitude and intensity of a collection of atoms is considered. Recall from equation 1.15 that the scattering from a number of identical dimensionless points can be expressed by:

$$A(\boldsymbol{q}) = |\boldsymbol{A_0}| b \sum_{i=1}^{N} e^{-i r_i \cdot \boldsymbol{q}}$$

---

[3] Unit analysis of equation 1.28: $\rho(r)$ is in units of Å^{-3}. Because we are in the spherical polar coordinate system, r has units of Å, and dr has units of Å. Thus, A(q) is dimensionless.

And, the scattering from a uniformly dense object expressed by the scattering length density distribution is given by equation 1.16:

$$A(\boldsymbol{q}) = A_0 \int_V \rho(\boldsymbol{r}) e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}\, dr$$

Now, b, the scattering efficiency, is a function of $\rho(\boldsymbol{r})$, and $e^{-i\boldsymbol{q}\cdot\boldsymbol{r}}$ in this equation relates the scattering of all vectors, **r**, within particle volume V in $\rho(\boldsymbol{r})$ to the corresponding scattering vector, **q**. This term in equation 1.16 can be redefined as:

$$A_V(\boldsymbol{q_V}) = A_0 \int_V \rho(\boldsymbol{r_V}) e^{-i\boldsymbol{q_V}\cdot\boldsymbol{r_V}}\, dr \qquad 1.29$$

to indicate that the scattering results from a particular scattering length density distribution in scattering volume V. In order to consider scattering from a number of atoms with defined scattering length density distribution, equation 1.29 and 1.15 can be combined(Glatter and Kratky, 1982):

$$A(\boldsymbol{q}) = |\boldsymbol{A_0}| \sum_{i=1}^{N} \left(\int_{Vi} \rho(\boldsymbol{r_{Vi}}) e^{-i\boldsymbol{q_{Vi}}\cdot\boldsymbol{r_{Vi}}}\, dr \right) e^{-i\boldsymbol{r_i}\cdot\boldsymbol{q}} \qquad 1.30$$

Since equations for a collection of *atoms* are being derived, $\int_{Vi} \rho(\boldsymbol{r_{Vi}}) e^{-i\boldsymbol{q_{Vi}}\cdot\boldsymbol{r_{Vi}}}\, dr$ is the atomic scattering factor for each atom, $f(\boldsymbol{q})$, (Drenth and Mesters, 2007) and can be substituted into equation 1.30:

$$f(\boldsymbol{q}) = \int_{Vi} \rho(\boldsymbol{r_{Vi}}) e^{-i\boldsymbol{q_{Vi}}\cdot\boldsymbol{r_{Vi}}}\, dr \qquad 1.31$$

$$A(\boldsymbol{q}) = |\boldsymbol{A_0}| \sum_{i=1}^{N} f(\boldsymbol{q}) e^{-i\boldsymbol{r_i}\cdot\boldsymbol{q}} \qquad 1.32$$[4]

---

[4] In crystallography, $A(\boldsymbol{q})$ in equation 1.32 is called the structure factor because it depends on the arrangement of the atoms in the molecule. This definition is slightly different than the “structure factor” definition used in small-angle x-ray scattering – see below.

This derivation highlights the fractal nature of scattering. The scattering amplitude of a molecule is dependent on the spatial arrangement of its monomer units. The monomer units of the molecule can be subdivided into smaller monomers and described by their spatial arrangement (Roe, 2000). The fractal nature of scattering is apparent until the monomer units are atoms. When the monomer units are atoms, the atomic form factor for each atom is defined as $f(\boldsymbol{q})$ and no further subdivision is possible.

The scattering intensity of a polyatomic molecule is given in equation 1.19:

$$I(q) = A(\boldsymbol{q}) \cdot A^*(\boldsymbol{q})$$

Therefore, substituting equation 1.32 in for $A(\boldsymbol{q})$:

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} f_i(\boldsymbol{q}) f_j(\boldsymbol{q})\, e^{-i\boldsymbol{q}\cdot\boldsymbol{r_{ij}}} \qquad 1.33$$

When $i = j$,

$$f_i(\boldsymbol{q}) f_j(\boldsymbol{q})\, e^{-i\boldsymbol{q}\cdot\boldsymbol{r_{ij}}} = {f_i}^2(\boldsymbol{q})$$

Since scattering is an orientational average of all scattering vectors in the molecule, when $i \neq j$,

$$I(q) = \sum_{i=1}^{N} \sum_{j=1}^{N} \langle f_i(q) \rangle \langle f_j(q) \rangle \frac{\sin(qr_{ij})}{qr_{ij}} \qquad 1.34$$

$\frac{\sin(qr_{ij})}{qr_{ij}}$ in equation 1.34 describes the orientationally averaged spatial relationship between atoms in a molecule. This term can also be applied more broadly to describe the orientationally averaged spatial arrangement of monomers within a polymer. Used in this way, $\frac{\sin(qr_{ij})}{qr_{ij}}$ is described as the structure factor.

Recall that f(q) is also orientationally averaged, so

$$\langle f(q) \rangle = 4\pi \int_{r=0}^{\infty} \rho(r) r^2 \frac{\sin qr}{qr} dr$$

describes the orientationally averaged scattering length density distribution for a particle with a given volume and scattering length density distribution. Therefore, I(q) in equation 1.34 is the orientationally averaged scattering from a molecule in solution (Glatter and Kratky, 1982).

1.6.6 A final note about notation

In section 1.6.6 the atomic scattering factor and the structure factor for a group of atoms have been defined. These definitions can be extended to apply to all monomer units where the scattering length density distribution is defined by $\rho(\boldsymbol{r})$. The form factor, F($\mathbf{q}$), describes the shape of the monomer electron density distribution in a polymer and the structure factor, S($\mathbf{q}$), describes the spatial arrangement of monomers within a polymer. More generally, the structure factor, S(q) is usually taken to be the orientational average of $e^{-i\boldsymbol{q}\cdot\boldsymbol{r}_{ij}}$ for all monomers i and j. Together, the form factor and structure factor fully describe the scattering from a molecule, and will be used in the modeling approaches described in the following chapters.

2. Modeling methods used in SAXS data analysis

2.1 Introduction

The interpretation of small-angle scattering data relies on the generation of structural models from the experimental scattering data. A major challenge in the structural modeling of SAXS data is determining which modeling methods are most appropriate (Putnam et al., 2007). In this chapter, three approaches that are often used to model proteins from SAXS data are described (Figure 8). These approaches are polymer physics based modeling (section 2.2), rigid body modeling (section 2.3) and ensemble based modeling (section 2.4). The purpose of this chapter is not to exhaustively review all modeling methods or programs[1], but rather to provide an overview of the most common methods, and to address the advantages and limitations of each method. Following a discussion of the modeling approaches, a discussion of SAXS data analysis as it applies to the modeling of highly flexible proteins will occur.

---

[1] In this chapter, *Ab Initio* modeling of molecules using SAXS data is not discussed. This method is still widely used and the reader is directed to the primary literature: Glatter, O., and Kratky, O. (1982). Small angle x-ray scattering (London ; New York: Academic Press), Svergun, D.I., Feĭgin, L.A., and Taylor, G.W. (1987). Structure analysis by small-angle x-ray and neutron scattering (New York: Plenum Press). Reviews: Koch, M.H., Vachette, P., and Svergun, D.I. (2003). Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Quarterly reviews of biophysics *36*, 147-227, Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Ibid. *40*, 191-285.

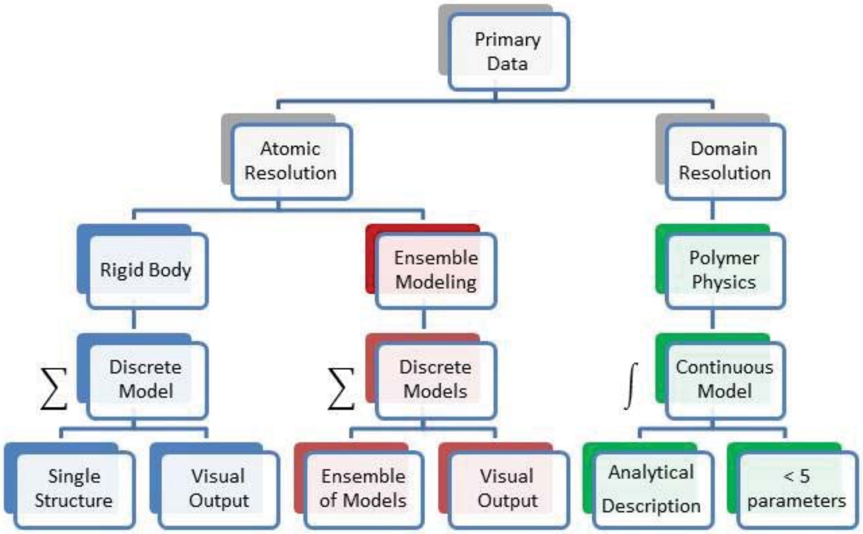

Figure 8: The three primary modeling approaches used to determine molecular structure from small-angle scattering data.

2.2 Polymer physics methods

2.2.1 Introduction

The simplest and least parameterized method of modeling the scattering behavior of macromolecules is found in the material science and polymer physics literature (Roe, 2000). Hence, this will be referred to as the polymer physics method. This method consists of describing the scattering length density of simple shapes (spheres, cylinders, ellipsoids, etc.) or flexible chains. Modeling complex data using simple shapes has a long history in the small-angle scattering community. Decades before there was a atomistic model of the ribosome from x-ray crystallography studies, Moore and colleagues developed a complete three-dimensional model of the ribosome using simple spheres to represent the various subunits (Engelman and Moore, 1972; Moore et al., 1986) .

A description of the scattering length density of simple shapes leads to a mathematical derivation of a scattering function. A model based on polymer physics theory assumes that a macromolecule is composed of uniformly dense simple geometrical objects and that the spatial relationship of these objects can be described analytically. The scattering function for a polymer model generally consists of five or fewer parameters.

2.2.2 Theory and principles

2.2.2.1 Scattering from simple geometric objects

A particle's scattering intensity can be approximated over all reciprocal space if the particle can be modeled as a collection of simple shapes. Recall that equation 1.28 describes the orientationally averaged scattering amplitude as:

$$\langle A(q)\rangle = 4\pi \int_{r=0}^{\infty} \rho(r) r^2 \frac{\sin qr}{qr} dr$$

Thus, if a function for the particle's scattering length density distribution, $\rho(r)$, is known, I(q) can be calculated.

The scattering function that is simplest to derive from equation 1.28 is that of a sphere (Guinier and Fournet, 1955). This is because a sphere has spherical symmetry and therefore does not require orientational averaging. The scattering length density distribution for a sphere of radius R with uniform density, $\rho_0 = 1$, is:

$$\rho(r) = \begin{cases} \rho_0, & r \leq R \\ 0, & r \geq R \end{cases} \qquad 2.1$$

This is depicted graphically in Figure 9A and 9B. Substituting equation 2.1 into equation 1.28, we obtain (Guinier and Fournet, 1955):

$$\langle A(q) \rangle = 4\pi \int_{r=0}^{R} \rho_0 r^2 \frac{\sin qr}{qr} dr \qquad 2.2$$

Since ρ_0 is a constant it can be removed from the integral:

$$\langle A(q) \rangle = \frac{4\pi}{q} \rho_0 \int_{r=0}^{R} r \sin(qr)\, dr \qquad 2.3$$

Integration of equation 2.3 by parts yields:

$$\langle A(q) \rangle = \frac{4\pi}{q} \rho_0 \left(\frac{-R\cos(qR)}{q} + \left[\frac{\sin(qr)}{q}\right]_{0}^{R} \right)$$

$$\langle A(q) \rangle = \frac{4\pi}{q} \rho_0 \left(\frac{-R\cos(qR)}{q} + \frac{\sin(qR)}{q^2} \right)$$

$$\langle A(q) \rangle = \frac{4\pi}{q^3} \rho_0 (-qR\cos(qR) + \sin(qR))$$

$$\langle A(q) \rangle = \frac{4}{3}\pi R^3 \rho_0 \frac{3(\sin(qR) - \mathrm{qR}\cos(qR))}{(qR)^3} \qquad 2.4$$

Equation 2.4 is the orientationally averaged scattering amplitude of a sphere. Since $\frac{4}{3}\pi R^3$ is the volume of a sphere with radius R,

$$\langle A(q) \rangle = V_s \rho_0 \frac{3(\sin(qR) - \mathrm{qR}\cos(qR)}{(qR)^3} \qquad 2.5$$

where $V_s = \frac{4}{3}\pi R^3$. Equation 2.5 emphasizes that a volume component will be present in all scattering functions of simple uniformly dense shapes[2] (Glatter and Kratky, 1982).

Because the intensity of scattering is (equation 1.19):

$$I(q) = A(\boldsymbol{q}) \cdot A^*(\boldsymbol{q})$$

or, for orientationally averaged amplitudes:

$$I(q) = \langle A(q) \rangle^2$$

then, for a sphere the scattering intensity as a function of q is:

$$I(q) = \rho_0^2 {V_s}^2 F(q)^2 \qquad 2.6$$

where $F(q)$ is the *form factor* of the sphere:

$$F(q) = \frac{3(\sin(qR) - \mathrm{qR}\cos(qR)}{(qR)^2} \qquad 2.7$$

The scattering intensity, I(q), for a single uniformly dense sphere of radius R is plotted in Figure 9C. The scattering functions of other simple shapes (e.g. cylinder (Glatter and Kratky, 1982), thin rod, and disk (Roe, 2000)) are plotted in Figure 9D.

---

[2] Dimensional analysis of equation 2.6: $V_s$is in units of Å^3, ρ_0 – like $\rho(r)$ – is in units of Å^{-3}, so A(q) and I(q) are dimensionless.

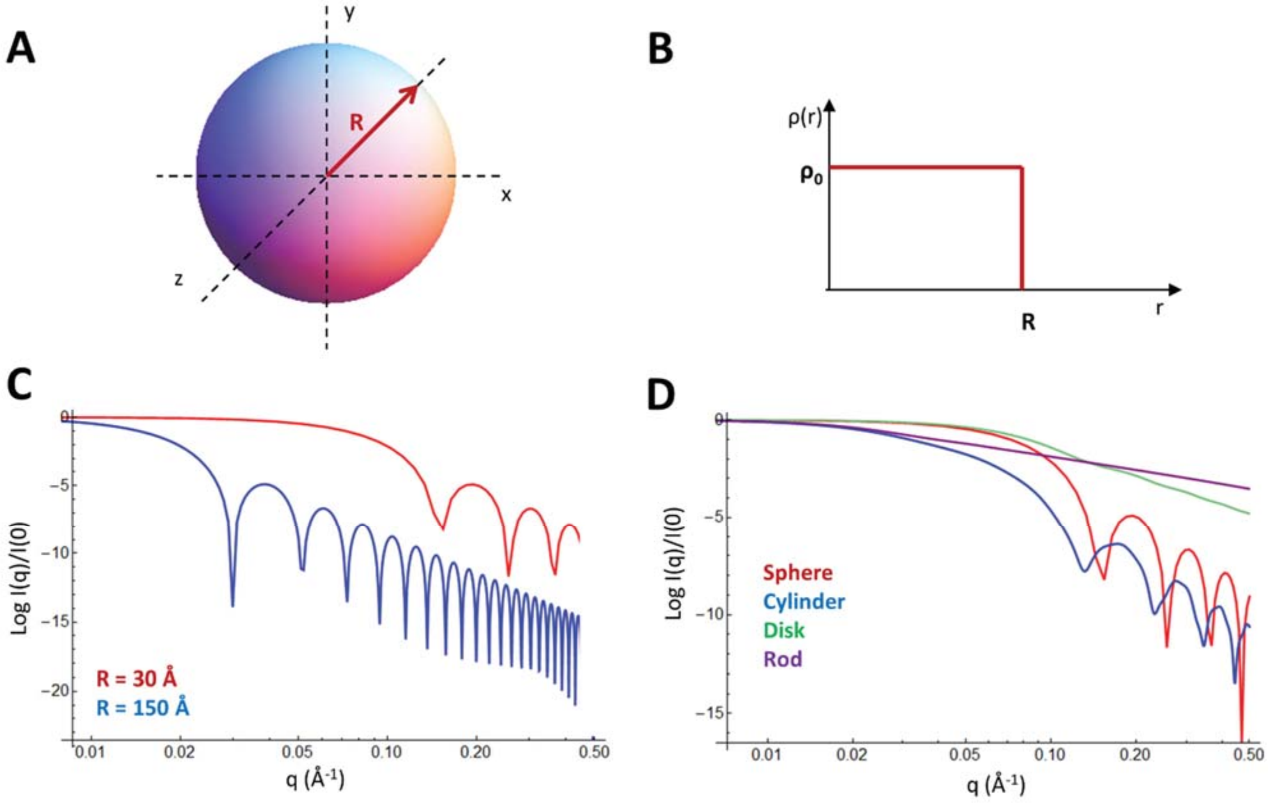

Figure 9: The form factors for simple objects. A) A sphere of radius R. B) The scattering length density distribution for a sphere of radius R. C) The form factors of two spheres of different radii: 30 Å (red) and 150 Å (blue). Notice that because of the inverse relationship between real and reciprocal space, the oscillations in scattering function of a sphere of 30 Å are larger than that of a sphere of 150 Å. D) The form factors for a sphere of radius 30 Å (red), A cylinder of radius 30 Å and length 200 Å (blue), a disk of radius 30 Å and thickness 2 Å (green) and a rod of radius 1 Å and length 200 Å (purple).

Similarly, there are analytical scattering functions for flexible polymers. Several excellent articles and books derive scattering functions for freely jointed chains, freely rotating chains, worm-like chains, etc. (Burchard and Kajiwara, 1970; Pedersen et al., 1996; Roe, 2000). It is important to note that in most cases the monomer unit for polymer chains is considered to be “nearly” dimensionless. In other words, all of the scattering from a monomer is concentrated at the center of the monomer in an infinitesimally small volume.

2.2.2.2 Observations of the scattering profile can help decide what simple shape model to use.

When equation 2.6 is plotted as log I(q) vs. log q, the repetitive nature of this function is emphasized. This feature results from the Fourier transform of the scattering length density distribution and is a common feature of all scattering functions of simple shapes. The nature of scattering from a uniformly dense sphere results in decaying oscillations in the scattering profile. Because scattering is the reciprocal space transform of real-space, the scattering function of a small sphere has fewer oscillations over the same scattering angles than a large sphere (Figure 9C). The size of a particle is a model-free parameter determined by observing the major features in the scattering profile. For a more comprehensive description of the relationship between the size of an object and its scattering profile refer to Roe, 2000.

The "power-law relationship[4]" between I and q is another model-free parameter determined from scattering data (Glatter and Kratky, 1982). This power law relationship states that the scattering intensity decays exponentially as the scattering angle increases and is asymptotic at large scattering angles:

$$I(q) \sim q^{-P} \tag{2.7}$$

where P is called the "Porod exponent." The Porod exponent is proportional to the ratio of a molecule's volume to its surface area. For a sphere, P = 4; in other words, it has a large volume : surface area ratio (Rambo and Tainer, 2011). The power law relationship is general for all simple shapes, and P is dependent on the compactness and dimension of the particle. For example, P=2 for infinitely thin disks (two-dimensional), and P=1 for infinitely thin rods (one-dimensional) (Figure 10A). Likewise, for flexible polymers, the value of P ranges from $\frac{5}{3}$ for "swollen Gaussian coils" to 3 for mass and surface fractals (Roe, 2000)[5].

Complex macromolecules exhibit different power-law behaviors at different scattering angles (Roe, 2000). For example, the intensity of a semi-flexible worm-like chain decays as $I(q) \sim q^{-2}$ at small scattering angles, but decays as $I(q) \sim q^{-1}$ at higher scattering angles.The transition from P=2 to P=1 is dependent on the flexibility of the

---

[4] The power law is also referred to as the Porod law in some books and articles

[5] Refer to Glatter and Kratky, 1982 for a good discussion of power law behaviors of different shaped objects.

chain (Figure 10B) (Pedersen, 1996)[6] . Generally, the first step in polymer modeling is to determine the power-law behavior in the experimental SAXS data. This observation assists in choosing the appropriate polymer model for the data.

---

[6] A classic work on scattering functions of worm-like chains is Burchard and Kajiwara, 1970.

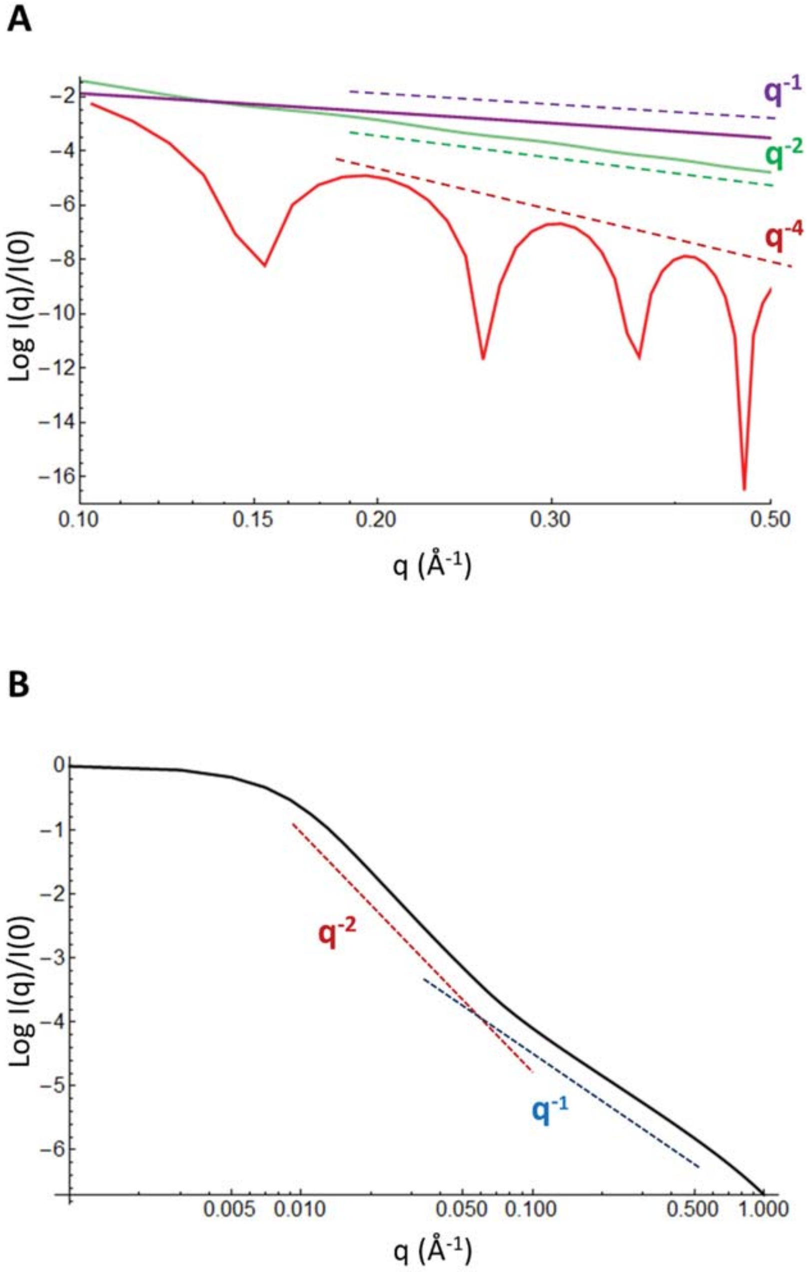

Figure 10: The power-law relationship between scattering intensity and scattering angle. A) At intermediate scattering angles ($0.15 < q < 0.5$ Å^{-1}) the scattering profile of simple shapes decays according to their volume : surface area ratio. Red: sphere, green: disk, purple: rod. B) The power-law relationship in a worm-like chain.

At low scattering angles, the intensity of scattering decays as q^{-2} and at high scattering angles it decays as q^{-1}.

2.2.2.3 Scattering from multi-phase particles

The scattering functions of polyatomic molecules were derived in section 1.6.6. Equation 1.34 was derived for all atoms in a molecule. Polymer physics modeling does not describe the shape of a polymer at the atomic level, it only describes the global shape of the particle. However, the shape of the particle does not necessarily need to be modeled as a single phase. For more precision, multi-phase models are often used to describe the scattering of particles. For a two phase system, a form factor is used to describe the scattering from a simple shape. The monomers within each phase are identical. Then, structure factors describing the spatial arrangement of the monomers within a phase can be derived. Finally, a cross-term must be derived that describes the scattering interference between monomers of two different phases. A simple model is described below.

Consider a polymer composed of two spherical monomers and two rods. Equation 1.34 defines the intensity of scattering for a polyatomic molecule as:

$$\langle I(q)\rangle = \sum_{i=1}^{N}\sum_{j=1}^{N}\langle f_i(q)\rangle\langle f_j(q)\rangle \frac{\sin(qr_{ij})}{qr_{ij}} \qquad 2.8$$

The set of monomers within this polymer is (sphere 1, sphere 2, rod 1, rod 2). If sphere 1 is identical to sphere 2 and rod 1 is identical to rod 2, then there are form factors, $(F_s(q), F_r(q))$ that describe the scattering of each type of monomer, where s denotes the form factor of the sphere and r denotes the form factor of the rod. The set of structure factors, used to describe the spatial arrangement of the monomers, is $(S_s(q), S_r(q), S_{sr}(q))$, where s denotes the spatial arrangement of the spheres – the scattering interference between spheres, r denotes the spatial arrangement of the rods – the scattering interference between rods, and sr denotes the cross-term – the scattering interference between rods and spheres. It is important to note that expressions $S_s(q), S_r(q),$ and $S_{sr}(q)$ are orientationally averaged structure factors that replace the expression $\frac{\sin(qr_{ij})}{qr_{ij}}$ in equation 1.34. The scattering intensity of this polymer is (Glatter and Kratky, 1982):

$$\mathrm{I(q)} = \sum_{i=1}^{2}\sum_{j=1}^{2} F_i(q)F_j\,(q)S_{ij}(q) \tag{2.9}$$

$$\mathrm{I(q)} = 2\mathrm{F}_s^2(\mathrm{q})\,\mathrm{S}_s(\mathrm{q}) + 4\mathrm{F}_s(\mathrm{q})\,\mathrm{F}_r(\mathrm{q})\,\mathrm{S}_{sr}(\mathrm{q}) + 2\mathrm{F}_r^2(\mathrm{q})\mathrm{S}_r(\mathrm{q}) \tag{2.10}$$

The scattering function for polymers with more than one phase are derived using equation 2.10, given the analytical functions for the form and structure factors of each phase, as well as the cross-term.

2.2.3 Software/fitting methods

The polymer physics literature is replete with scattering functions for a variety of polymers. There are several software packages that enable the fitting of experimental scattering data to these polymer models. The two most common software packages are SASView (www.sasview.org), and NCNR_SANS (Kline, 2006).

NCNR_SANS fits experimental scattering data to polymer models using a non-linear least squares fitting algorithm implemented in IgorPro. The regularly updated model function library contains all the scattering functions published in the scientific literature – both form factors and structure factors.

SASView is essentially an implementation of NCNR_SANS in a Python computing environment. It contains all the model functions found in NCNR_SANS, but relies on non-linear least squares fitting as implemented in SciPy (a Python module) to fit experimental data. SASView also allows for the global fitting of a series of scattering curves to the same model. This feature is important for fitting both SANS (small-angle neutron scattering) contrast variation experiments, as well as SAXS data from polymers with varying numbers of monomers.

2.2.4 Limitations of the polymer physics methods

A primary assumption in polymer physics modeling is that monomer units possess uniform electron density. Thus, the application of this method to biological macromolecules is limited to low resolution studies. Although polymer physics models can describe the overall dimensions and low resolution shape of a macromolecule they may not adequately describe the scattering at the resolution level of protein secondary structure. Therefore, if details about the structure of protein domains are the goal, a method that takes into consideration the non-uniformity of protein secondary structure may be better for interpreting SAXS data.

Another limitation of the polymer physics model is that, for flexible polymers, care must be taken to account for the excluded volume effects of real polymers. Many scattering functions for polymers with a very large monomer number are based on the “random flight” flexible polymer model. This model allows for overlap between hard sphere monomers. In other words, two monomers can inhabit the same volume. While this approximation has been shown to be valid for polymers consisting of hundreds or thousands on monomers, it may not be valid for short polymers. Any analysis of experimental scattering data using a polymer physics approach must take into consideration these two limitations. Depending on the hypothesis being tested, higher resolution modeling might be more appropriate, depending on the information content of the SAXS data.

2.3 Rigid body modeling

2.3.1 Introduction

When an atomistic model of a protein exists, the theoretical scattering profile of the atomistic model can be compared to the experimental scattering profile. Often this approach is useful for distinguishing between two or more protein structures obtained from other experimental techniques (x-ray crystallography or NMR). This approach can also be used when looking at complexes of macromolecules. Comparing the SAXS profiles predicted for different arrangements of the component structures allows for the selection of an atomistic model of the solution structure of the complex (Putnam et al., 2007). This modeling method relies on an accurate and appropriate theoretical scattering profile determined from component atomistic models.

2.3.2 Theory and principles

2.3.2.1 The Debye sum – reconstructing scattering profiles from atomistic models is an O(N^2) problem

Reconstructing theoretical scattering profiles from atomistic models can be computationally challenging. Several approximations and simplifications have been proposed to overcome this challenging problem. The discrete form of equation 1.34 gives the scattering profile for a polyatomic model.

$$I(q) = \sum_{i=1}^{N}\sum_{j=1}^{N}\langle f_i(q)\rangle\langle f_j(q)\rangle\frac{\sin(qr_{ij})}{qr_{ij}}$$

for all pairs of atoms in a molecule. This is commonly called the "Debye sum." $\langle f_i(q)\rangle$ and $\langle f_j(q)\rangle$ are the atomic form factors (Glatter and Kratky, 1982). $\frac{\sin(qr_{ij})}{qr_{ij}}$ describes their spatial relationship. The complexity of this computation is on the order of N^2, (O(N^2), where N is the number of atoms, for each q evaluated because all pairs of atoms must be considered in the double sum. For large proteins, the number of atoms is very large ($N \geq 10^5$) so the computational cost of the calculation can be prohibitive (Gumerov et al., 2012). While some rigid body modeling algorithms explicitly compute the Debye sum for all atoms in the model and overcome the computational complexity by parallelization of the algorithm (Gumerov et al., 2012), other algorithms seek to simplify the Debye sum by a series of approximations in order to speed up the calculation[7].

The first approximation of the Debye sum is the multipole expansion of $e^{-i\boldsymbol{q}\cdot\boldsymbol{r}_i}$ proposed by Sturhman and implemented in CRYSOL (Svergun et al., 1995). This substitution expands the pre-orientationally averaged sum into a series of spherical harmonics. The spherical harmonics are then orientationally averaged to approximate the scattering profile of the protein. It is important to note that the multipole expansion is not an infinite series of spherical harmonics. Instead the series expansion is truncated to a small number of harmonics, p (the CRYSOL default value is 15 harmonics), so that the computational complexity of computing the sum is reduced to $O(p^2N)$. However, this approximation may not accurately calculate the theoretical scattering profile from an atomistic model, particularly for small p values. Higher p values result in a better approximation, but increase the computational time of the algorithm.

A related method for simplifying the Debye sum is to use a three-dimensional Zernike polynomial expansion instead of spherical harmonic expansion in the

---

[7] Refer to Table 1 for a summary of the types of techniques used to overcome the computational complexity of the Debye sum

approximation (Liu et al., 2012). This method is similar to the multipole expansion and results in a similar reduction of computational complexity.

Another method calculates the pair-distance distribution function, P(r)[8], of the atomistic model in order to simplify the Debye sum. The P(r) function is a histogram of all the interatomic distances within the model. This histogram is separated into a series of coarse-grained bins (D. Walther, 2000) and the Debye sum is approximated using the binned interatomic distances in the P(r) function. This method is computationally faster than the direct Debye sum. However, the computational time is still limited by the number of bins, and by the generation of the initial P(r) function, which still has $O(N^2)$ complexity. Additionally, this method results in an approximation of the theoretical scattering profile: the "resolution" of the resultant scattering profile is limited by the number of bins used in the P(r) function (Gumerov et al., 2012).

A coarse-grained technique for calculating the theoretical scattering profile from an atomistic model is to represent an entire protein residue with a single form factor centered on the Cα carbon instead of using a calculated form factor for each individual atom (Yang et al., 2009). This effectively reduces the $O(N^2)$ calculation to $O(R^2)$, where R is the number of residues. However, while this approximation decreases computational time, it does not account for rotamer differences in the amino acids.

---

[8] P(r) is defined as $\rho(r)r^2$.

2.3.2.2 Accounting for the hydration shell around a protein

All of the scattering equations derived to this point are for a molecule in a vacuum. In practice, the scattering from a molecule or a collection of molecules is observed while in an aqueous solution. The true scattering data is therefore obtained by subtracting the scattering profile of a perfectly matched buffer solution from the protein + buffer experimental SAXS data (Roe, 2000):

$$I_{protein}(q) = I_{protein+buffer}(q) - I_{buffer}(q)$$

This approach works in polymer physics modeling because the scattering object is assumed to have a constant electron density. However, if we model the scattering of a protein in solution using an atomistic model, then the hydration shell surrounding the protein must also be accounted for in the model.

Various methods have accounted for the hydration shell by assuming that the scattering length density distribution in the hydration shell is different from that of either the bulk solvent or the protein. Thus, these methods introduce one or more adjustable parameters into the calculation of the theoretical scattering profile to account for the contrast difference.

The method implemented in CRYSOL models the hydration layer using a continuous envelope around the protein model (Svergun et al., 1995). The thickness and scattering efficiency of this border layer are adjustable parameters that are optimized when fitting the theoretical scattering profile to experimental data. One limitation of this method is that if the protein has a hollow core or is ring-shaped, the hydration layer of the inside ring is not considered.

The method used in FOXS represents the hydration layer using a series of spherical dummy atoms that “decorate” the surface of the protein (Schneidman-Duhovny et al., 2010). There are three adjustable parameters in this model: two coefficients that model the excluded solvent and bound surface water, and a term for the solvent accessibility of the surface atom. The parameters that model the excluded solvent and bound surface water determine the size of the dummy atoms. The solvent accessibility, calculated from the atomistic model, is used to determine where the dummy atoms are placed in the hydration layer surrounding the protein. In other words, the solvent accessibility determines the number of dummy atoms. One limitation of this method is that it can result in a non-uniform density of the hydration layer because it allows the dummy atoms to overlap and there are empty spaces between spherical dummy atoms. Portions of the density are “counted twice” while other portions of an assumed constant hydration layer are not represented at all in this approximation.

The method used in SASTBX to model the hydration layer around atoms is called the modified cube approach (Liu et al., 2012). This approach uses the Zernike polynomial expansion to approximate the Debye sum from the atomistic model, and likewise accounts for the hydration layer within the voxelization procedure that converts the atomistic model into a series of cubic voxels. The hydration layer is then modeled as cubic voxels around the protein electron density. This method avoids the over- and under- counting of the hydration layer that occurs when using the dummy atom approach, but it does rely on an approximation of the total Debye sum.

Finally, in AXES, ORNL-SAS, and Fast-SAXS water molecules are explicitly added to the atomistic model using molecular dynamics simulations to place the water molecules along the solvent accessible surface area of the protein (Grishaev et al., 2010) (Gumerov et al., 2012),(Ravikumar et al., 2013).

2.3.3 Software

A summary of common methods for predicting the theoretical scattering profile is given in Table 1. In all cases except for the Hierarchical Algorithm, the computational complexity of calculating a theoretical scattering profile from atomistic models is decreased by either approximating the Debye sum or modeling an implicit, thus minimally parameterized, hydration layer. When explicit water models are used in these methods to calculate the hydration layer, the location of the water molecules is determined by some form of energy function. It is not always clear what these energy functions are based on, or whether they are physically realistic. Thus even explicit water modeling may be considered an approximation.

Table 1: Methods for theoretical SAXS profile calculation

Program	Calculation of the Debye sum	Calculation of the hydration layer
CRYSOL (Svergun et al., 1995)	Spherical harmonics approximation	Implicit solvent envelope
FOXS (Schneidman-Duhovny et al., 2010)	Direct calculation	Dummy atoms
AXES (Grishaev et al., 2010)	Spherical harmonics approximation	Explicit
SASTBX (Liu et al., 2012)	Zernike polynomial expansion	Modified cube approach
Fast-SAXS (Ravikumar et al., 2013)	Coarse-grain residue approximation	Explicit
ORNL_SAS (Tjioe, 2007)	Binning	Explicit
Hierarchical algorithm (Gumerov et al., 2012)	Direct calculation	Explicit

It is interesting to compare the theoretical SAXS profiles calculated by each of these methods when there is no experimental SAXS data to constrain the results[9]. Figure 11 shows the theoretical scattering profiles calculated for two atomistic models: lysozyme (pdb: 3A8Z) and the antibody, IgG2 (pdb: 1IGT). Lysozyme is often used as a standard in SAXS experiments and is a globular molecule. IgG2 is spherically asymmetric.

There is significant agreement in the very low q region between the various theoretical scattering curves calculated for both lysozyme and IgG2 (Figure 11A,C). This

---

[9] Neither ORNL_SAS or the Hierarchical Algorithm are publically available, so no theoretical SAXS generated by these methods are presented.

region is representative of the global structure of the molecule. However, there is no agreement between the unconstrained theoretical scattering curves at high q for either lysozyme or IgG2. This region of the scattering curve is dependent on the intra-domain structure of the molecule and the hydration layer surrounding the molecule. Thus, variations in how the Debye sum is calculated and in how the hydration layer is approximated contribute to the lack of agreement between methods.

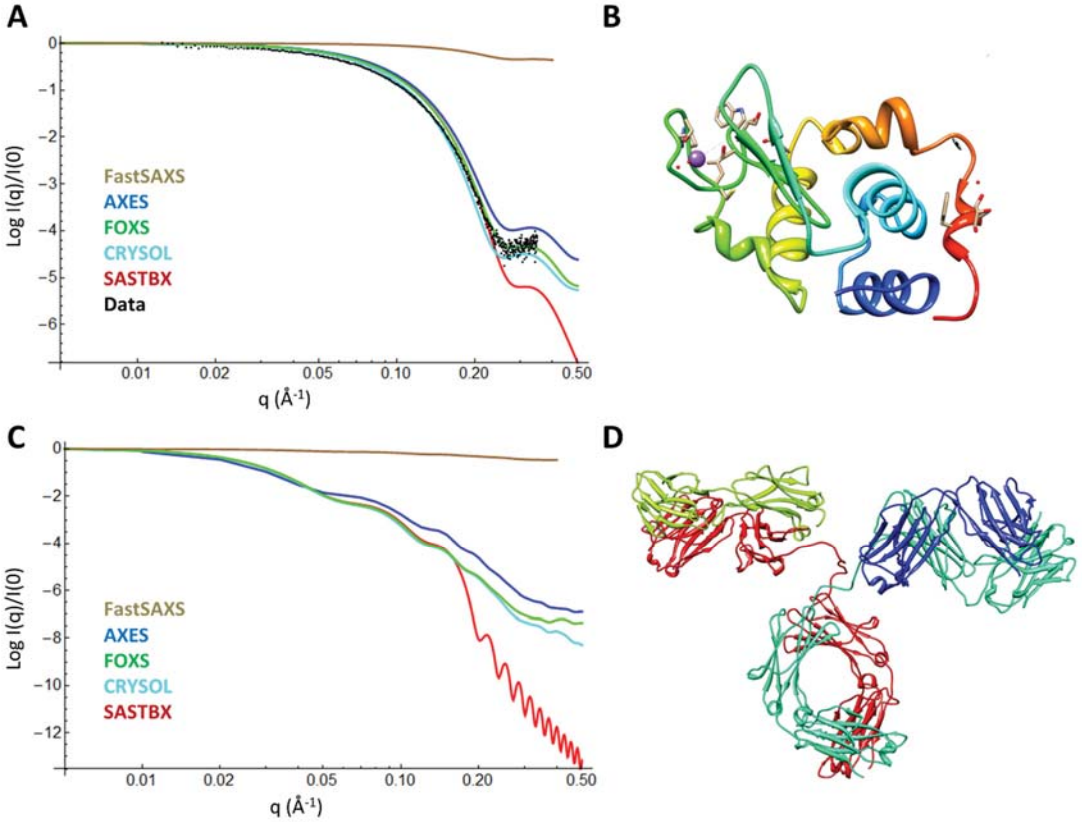

Figure 11: Comparison of theoretical scattering curves of two proteins: lysozyme (pdb: 3A8Z) and IgG2 (pdb: 1IGT). A) Theoretical scattering curves calculated by 5 programs for lysozyme, using 3A8Z as the input molecule. Black: Experimental data obtained from the Bio-Isis database (bid: LYSOZP) B) Structure of

lysozyme. C) Theoretical scattering curves calculated by 5 programs for IgG2, using 1IGT as the input molecule. D) Structure of IgG2. Figures 4B and 4D are not to scale.

In addition, when the un-constrained theoretical scattering profiles calculated for lysozyme are compared to experimental data (Figure 11A), no single theoretical scattering curve fully predicts the experimental data over the entire q range. At low scattering angles, the SASTBX program best fits the experimental data. However, at high scattering angles this is no longer true. In this q-region, the data is best fit by the FOXS theoretical scattering profile.

2.3.4 Limitations of rigid body modeling

As seen in Figure 11, one of the limitations of rigid body modeling is that the approximations used to generate a theoretical scattering profile from an atomistic model may not accurately reflect the scattering of the molecule in solution. At high q, there is no agreement in the theoretical scattering curves calculated by each of these programs, and it is impossible to determine which method may “best” reflect the scattering of the molecule in solution at all angles. This limitation is generally addressed by using the experimental data to constrain the theoretical profile. The adjustable parameters that account for the hydration layer are optimized against the experimental data. However, it is not always clear that these parameters reflect the physical properties of the molecule in solution; instead they may often be used as “fudge factors” to optimize the fit of the model to the data (Virtanen et al., 2010).

Another limitation of rigid body modeling is its assumption that a single atomistic model adequately describes the conformation of the molecule in solution. For rigid and compact proteins this assumption may generally be true, particularly when one considers the “low resolution” of SAXS data (Putnam et al., 2007). However, for flexible proteins, this assumption is may be incorrect. A flexible protein may adopt many conformations in solution, and this conformational complexity is reflected in the SAXS data.

Small-angle scattering data only reflects the global properties of the molecule in solution and detection of small-scale conformational heterogeneity (movement of side chains) is beyond the resolution limit of the experimental technique. However, small-angle scattering is very sensitive to large scale heterogeneity in the conformational ensemble of the molecule (Bernado et al., 2007). The scattering profile of a molecule is the population-weighted sum of the scattering profiles of all conformations in solution. Such conformational shifts are not captured by discrete atomistic models. When large inter-domain motions are suspected, ensemble-based modeling or polymer physics modeling may better reflect the conformational ensemble of the protein.

2.4 Ensemble-based modeling of the macromolecule

2.4.1 Introduction

Ensemble-based modeling compares the experimental SAXS data to theoretical aggregate scattering curves calculated from atomistic models by some ensemble-generating algorithm. This approach relies on the fact that the observed scattering of a molecule in solution is the population-weighted sum of the scattering from all conformers of the molecule (Hura et al., 2013).

Modeling the conformational ensemble generally involves two steps. The first step is to generate a large number of atomistic models using known structural constraints. This large parent ensemble is whittled down to a minimal ensemble – a subset of conformations from the parent ensemble whose theoretical scattering profiles are best able to fit the experimental scattering data. These two processes as well as some programs that perform this calculation are discussed in detail below.

2.4.2 Generation of the "parent ensemble" – conformational sampling

The first step in any type of ensemble modeling is to generate a series of conformers that ideally sample the sterically allowed conformational space of a molecule. This step requires some assumptions about the structure and rigidity of the molecule. Experiments such as hydrogen-deuterium exchange (Hernandez and LeMaster, 2009), NMR dynamics measurements (Bracken, 2001), and limited proteolysis (Fontana et al., 2004) can identify regions of flexibility and rigidity within the molecule.

Once the flexible regions of the molecule are identified, several different methods may be used to generate a parent ensemble of structures (usually 10,000 or more) that sufficiently samples conformational space (Bernado and Svergun, 2012). In all cases, rigid body modeling of the domains in a flexible protein is used to reduce the computational complexity of the molecular modeling. The computations used to generate the parent ensemble are applied only to those regions where flexibility is indicated.

The most basic method for generating the parent ensemble is that implemented in RanCH (Bernado et al., 2007) and Flexible-Meccano (Ozenne et al., 2012). In this method, backbone dihedral angles for each amino acid in the flexible region are sampled from a population-weighted Ramachandran distribution unique to each amino acid. Rigid body rotations and translations of the rigid domains are then performed around the flexible linkers. A hard sphere potential for each amino acid is applied, and conformations are rejected when steric clashes occur. In this method no other force-fields are applied to account for attractive or repulsive forces between amino acids.

Another method, similar to the one above, is SASSIE (Curtis et al., 2012). In this method, the backbone dihedral angles are randomly chosen from a weighted Ramachandran distribution of allowed dihedrals. It is important to note that this is a single Ramachandran distribution, instead of a unique distribution for each amino acid. Once the atomistic model of the conformation is created using dihedral sampling, the model is energy minimized by CHARM 22 (Mackerell et al., 2004) to account for hard-sphere constraints and attractive and repulsive forces. Thus, this method uses a combination of geometric constraints and molecular dynamics (MD) force fields to generate each model in the parent ensemble.

A third set of methods for generating parent ensembles is the minimal molecular dynamics approaches implemented in BILBO-MD (Pelikan et al., 2009) and BSS-SAXS (Yang et al., 2010). In these methods, the initial model is subjected to a high-temperature simulated annealing algorithm to identify conformations with a local energy minimum. This MD simulation is only applied to the flexible regions of the protein. A rigid body rotation and translation is applied to the protein domains around the flexible regions. One unique aspect of the BILBO-MD algorithm is that force fields for electrostatic and Van der Waals interactions are excluded from the energy function. This simplification speeds up the calculation of the parent ensemble considerably. After generating the atomistic model, a second energy minimization is performed, this time with two other constraints: harmonic constraints on the protein eliminate any hard-sphere steric clashes, and a user-generated radius of gyration (R_g) constraint eliminates all models with an R_g outside particular boundaries determined from the experimental SAXS data. Thus, this is the only method reviewed here that uses the SAXS data as a constraint when generating the parent ensemble.

2.4.3 Selection of the “minimal ensemble”

Once the parent ensemble has been generated, it is general practice to select a minimal ensemble that describes the experimental scattering data. For flexible proteins, it is important to remember that this minimal ensemble may not be unique, but instead reflects the properties of the entire statistical conformation of the molecule. In general, minimal ensembles consist of 1-50 structures, but the size of a minimal ensemble varies depending on the method used to generate it.

One method of generating a minimal ensemble, implemented in the Ensemble Optimization Method (EOM) (Bernado et al., 2007) and the Minimal Ensemble Search (MES) (Pelikan et al., 2009), is to use a genetic algorithm. First, the theoretical scattering profiles for each conformer in the parent ensemble are calculated. Then, the scattering profiles in the parent ensemble are randomly divided into subsets. The subsets are subjected to random mixing between subsets, duplication of individual conformers, or deletion of conformers in the subset. The resulting aggregate SAXS profile from each subset is compared to the experimental data and the subset with the best fit-statistics to the experimental data is carried through to the next cycle of the genetic algorithm. Generally, 1000 or more cycles are performed. In this genetic algorithm, the mutations in the subsets during each cycle are not constrained by the experimental data. At the end of the computation, a weighted population of minimal conformers is obtained. In the EOM approach, usually 10-50 conformers are used to describe the conformational ensemble.

In the MES approach, 2-5 conformers are enumerated. In both cases, the authors emphasize that although the minimal ensembles are depictions of the conformational ensemble, the conformations in the minimal ensemble do not represent the only conformers in solution. Only the radius of gyration (R_g) and R_g distribution of the minimal ensemble should be considered when interpreting the results of the algorithm.

ENSEMBLE (Krzeminski et al., 2013) and EROS (Rozycki et al., 2011) use a different approach for generating a minimal ensemble from a parent ensemble. In this approach an ensemble with a maximum entropy weight distribution is selected. All the conformers in the parent ensemble are first clustered by a pair-wise comparison of the average root-mean-squared-distance between each conformer[10]. Then, an aggregate scattering curve for each cluster is calculated and compared to the experimental data. A maximum entropy and simulated annealing approach is used to determine the population of each cluster that in total best fits the experimental scattering curve. Thus, the output of the algorithm is a series of conformers and their respective weights that best fit the experimental scattering curve. In contrast to the EOM and MES methods above, the authors of EROS state that this method not only provides information about the global features of the conformational ensemble, but that the atomistic models fully

---

[10] It was noted that, for the test protein, 40% of all clusters contained a single conformer.

represent the solution conformations and can be used to explain specific chemical and biological processes.

2.4.4 Limitations of ensemble modeling

The major limitation of these methods is the ability to accurately predict the theoretical scattering curve of a molecule from a given atomistic model. As illustrated in Figure 3, this prediction is by no means fool-proof. When experimental SAXS data is not used to constrain the model, each method generates vastly different scattering profiles from the atomistic model. Thus, one has to question whether ensemble selection methods truly provide more information than distributional information of the global structural parameters, given how reliant they are on *ab initio* generation of theoretical scattering curves.

Another limitation of this approach is that the parent ensembles may significantly under-sample allowed conformational space (Berlin et al., 2013). For an unfolded protein or a flexible linker between two domains, Levinthal's Paradox (Levinthal, 1968) states that since a random coil peptide chain has a very large number of degrees of freedom, its number of possible conformations is astronomically large. For instance, if a two-domain protein is separated by 31 flexible residues, it has 30 peptide bonds and 60 different psi/phi bond angles. If each phi and psi angle is allowed to adopt three stable conformations, then number of possible conformations of the peptide is on the order of $3^{60} = 10^{28}$ conformations. Generally, ensemble based methods sample 10,000 conformations – an infinitesimally small fraction of the total number of possible conformations. Thus, care must be taken to assure that the whole of conformational space is explored and that the parent ensemble is truly a random sampling of conformational space.

Another limitation of ensemble modeling is verification of the robustness and uniqueness of the structural parameters of the minimal ensemble. As will be shown in Chapter 4, sometimes the solutions generated by the minimal ensemble enumeration methods are neither robust nor unique.

2.5 Modeling the structure of highly flexible macromolecules

2.5.1 Introduction

Rigid body and ensemble modeling approaches have been very successful in modeling the structures of rigid and minimally flexible proteins (Putnam et al., 2007). For a protein that exists in two conformations, open and closed, rigid body modeling has been able to distinguish between these two conformations, and ensemble modeling has been able to assign weights to the relative populations of these two conformations in different solution conditions (Petoukhov and Svergun, 2005). However, these techniques are not always successful in determining “structures” of highly flexible proteins (Receveur-Brechot and Durand, 2012) because of the limitations discussed above. In the following sections I will discuss some of the problems in modeling "structures" of highly flexible proteins from SAXS data, review methods for determining if a protein is flexible, and provide a method utilizing polymer physics and ensemble approaches to describe the statistical conformations of highly flexible proteins.

2.5.2 The information content of SAXS data

It has long been recognized that small-angle scattering data analysis is an ill-posed problem: the number of adjustable parameters used in SAXS analysis normally far exceeds the number of independent observables in a SAXS dataset (Gumerov et al., 2012; Rambo and Tainer, 2013). For rigid molecules, this lack of observables results from the lack of ordering of molecules in solution (Konig et al., 1993). Due to the orientational averaging of molecules in solution, it is challenging to model a 3D structure of a molecule from the 1D SAXS dataset (Vachette et al., 2003).

For flexible proteins, this problem is also compounded by the sheer number distinct conformations that may be present in solution in any given condition. Berlin and co-workers (Berlin et al., 2013) have recently suggested that for poly-ubiquitin 2, a two domain protein with a flexible inter-domain linker, the number of independent observables in a SAXS dataset may be as low as 3. If atomistic modeling is used to explain this type of SAXS data, a number of non-unique ensembles of models may fit the experimental data with the same accuracy. In these cases, interpretation of the SAXS data using atomistic models may result in over-parameterization of the model. More information about the structure of the protein results from the modeling technique used than can be inferred from the experimental data alone.

Despite these limitations, small-angle scattering analysis yields important information about the statistical conformations of highly flexible proteins and this information been confirmed by other experimental techniques (Stollar et al., 2012). Recognizing the importance of SAXS as a biophysical tool, I do not want to discourage its use in the study flexible molecules. Rather, I advocate careful consideration of the appropriateness of each modeling approach, given the experimental question being addressed and the information content of the scattering data. The number of parameters in the model must be matched to the information content of the SAXS data.

2.5.3 Indicators of flexibility in the scattering data?

The first step in modeling molecules based on SAXS data is to determine if the molecule is compact and globular, spherically asymmetric, or flexible. In the biological SAXS literature, there are two recognized “tests” for flexibility that can be performed with the primary SAXS data without the need for modeling (Rambo and Tainer, 2011; Receveur-Brechot and Durand, 2012). Both of these tests, the dimensionless Kratky plot and the Porod-Debye plot, are based on the power-law relationship between the scattering intensity and the scattering angle. Although these are widely accepted tests used to distinguish between rigid proteins and flexible proteins, using simple shape models I will show that these tests can distinguish between globularity and non-globularity but cannot distinguish between flexibility and spherical asymmetry.

2.5.3.1 The dimensionless Kratky plot

As seen in section 2.2.2.2, a log I vs. log q plot of the scattering data can yield important information about volume : surface area ratio of a particle at different length scales. This information can then be used to start to describe the "shape" of the particle. While the Porod exponent can be obtained from the log-log plot of the scattering data, it is difficult to determine where the transition occurs between one type of power-law relationship and another. Several transforms of the scattering data can help clarify where this transition occurs.

In the biological SAXS literature, the dimensionless Kratky plot has been promoted as a useful tool to distinguish between rigid and flexible biomolecules (Hammel, 2012; Mertens and Svergun, 2010; Pelikan et al., 2009; Williams et al., 2012). In this plot, the x-axis, q, is scaled by the molecule's R_g, and the y-axis is $\left(qR_g\right)^2 I(q)/I(0)$. The Kratky plot allows one to easily determine if a SAXS profile exhibits a $P \approx 4$ power-law relationship.

For compact molecules, where $P \approx 4$, there will be a peak at $qR_g \approx \sqrt{3}$ and $\left(qR_g\right)^2 I(q)/I(0) \approx 3e^{-1}$ (Durand et al., 2010; Rambo and Tainer, 2011). For a completely random coil, $P = {}^{5}/_{3}$, and on the Kratky plot this is indicated by a horizontal asymptote at $\left(qR_g\right)^2 I(q)/I(0) \approx 2.5$. Finally, for a rigid rod, the slope of the scattering intensity will increase linearly. Figure 12 shows the dimensionless Kratky plot for a sphere, cylinder, disk, rod, and random coil. Both the sphere and disk display a peak at $qR_g \approx \sqrt{3}$, indicating, by the above criteria, that the particle is rigid. However, the cylinder, which is very spherically asymmetric, does not have a peak at $qR_g \approx \sqrt{3}$. In addition, as mentioned above, the rod, which can be considered rigid, also does not display a peak at $R_g \approx \sqrt{3}$. Instead, $\left(qR_g\right)^2 I(q)/I(0)$ increases linearly with qR_g. Thus, for elongated spherically asymmetric molecules, it is difficult to distinguish between "flexibility" and spherical asymmetry. Characteristic features in the dimensionless Kratky plot can only indicate deviation from globularity *or* flexibility in conformational ensemble of the molecule. Nevertheless, this plot is a helpful first-assessment of non-globularity.

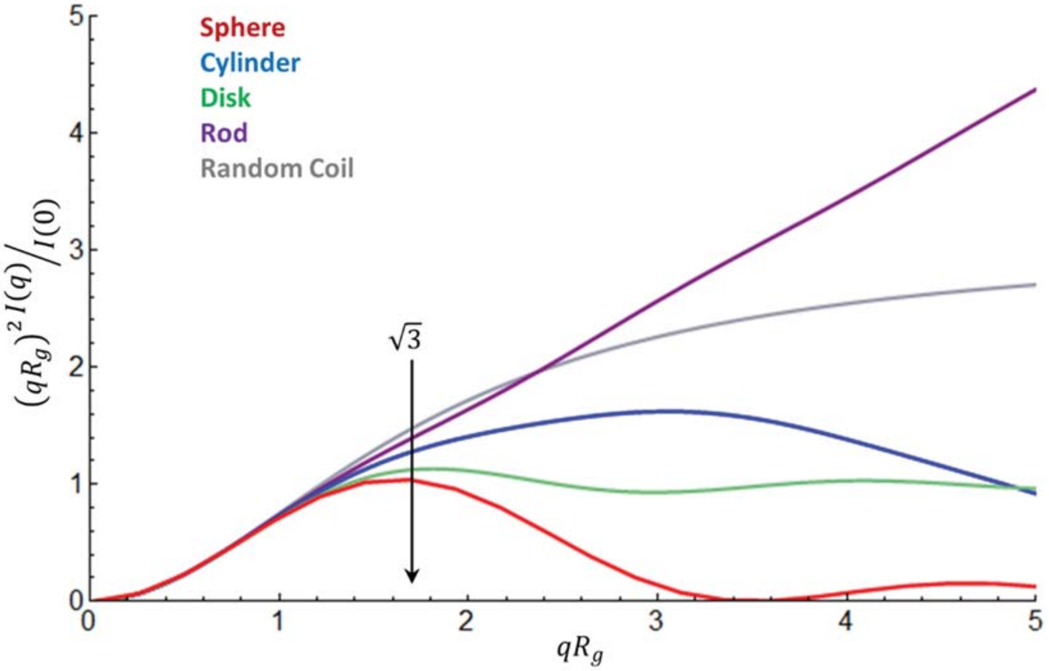

Figure 12: The dimensionless Kratky plot for a sphere (red), cylinder (blue), disk (green), rod (purple), and random coil (grey).

2.5.3.2 The Porod-Debye plot

Rambo and colleagues have suggested that the Porod-Debye plot may also be used to determine the flexibility of a protein (Hammel, 2012; Hura et al., 2013; Rambo and Tainer, 2011, 2013). In this plot, the scattering data is plotted as $q^4\, I(q)/I(0)$vs. q^4. The authors assert that at low scattering angles the scattering of a folded protein should decay as q^{-4}. Therefore, in the Porod-Debye plot folded proteins will display a horizontal plateau at low scattering angles. While this observation of a horizontal plateau is true for globular folded proteins, it may not be true for spherically asymmetric folded proteins. Figure 13 shows the Porod-Debye plot for basic shapes, meant to simulate folded globular and folded spherically asymmetric molecules. Both the sphere (Figure 13A) and cylinder (Figure 13B) Porod-Debye plots have a horizontal plateau, suggesting, by Rambo's criteria, that these are "folded" shapes. This is in contrast to the dimensionless Kratky plot, where the cylinder did not exhibit the characteristic features of globular shape. Interestingly, the Porod-Debye plot for the disk (Figure 13C) suggests that this shape is "not folded", in contrast to the dimensionless Kratky plot that exhibited the characteristic features of a globular shape. The Porod-Debye plot for the rod (Figure 13D) suggests that this shapes is "not folded." However, one can imagine that an elongated biomolecule, such as double-stranded DNA, might best be described as either a cylinder or a rod, and it is certainly a folded, homogenous biomolecule in solution. Thus, I suggest that while the Porod-Debye plot may be helpful in visualizing the globularity of a molecule, lack of a horizontal asymptote at low scattering angles does not indicate that the molecule is flexible, only that it may be spherically asymmetric. This plot, like the dimensionless Kratky plot, can only distinguish between globularity and non-globularity.

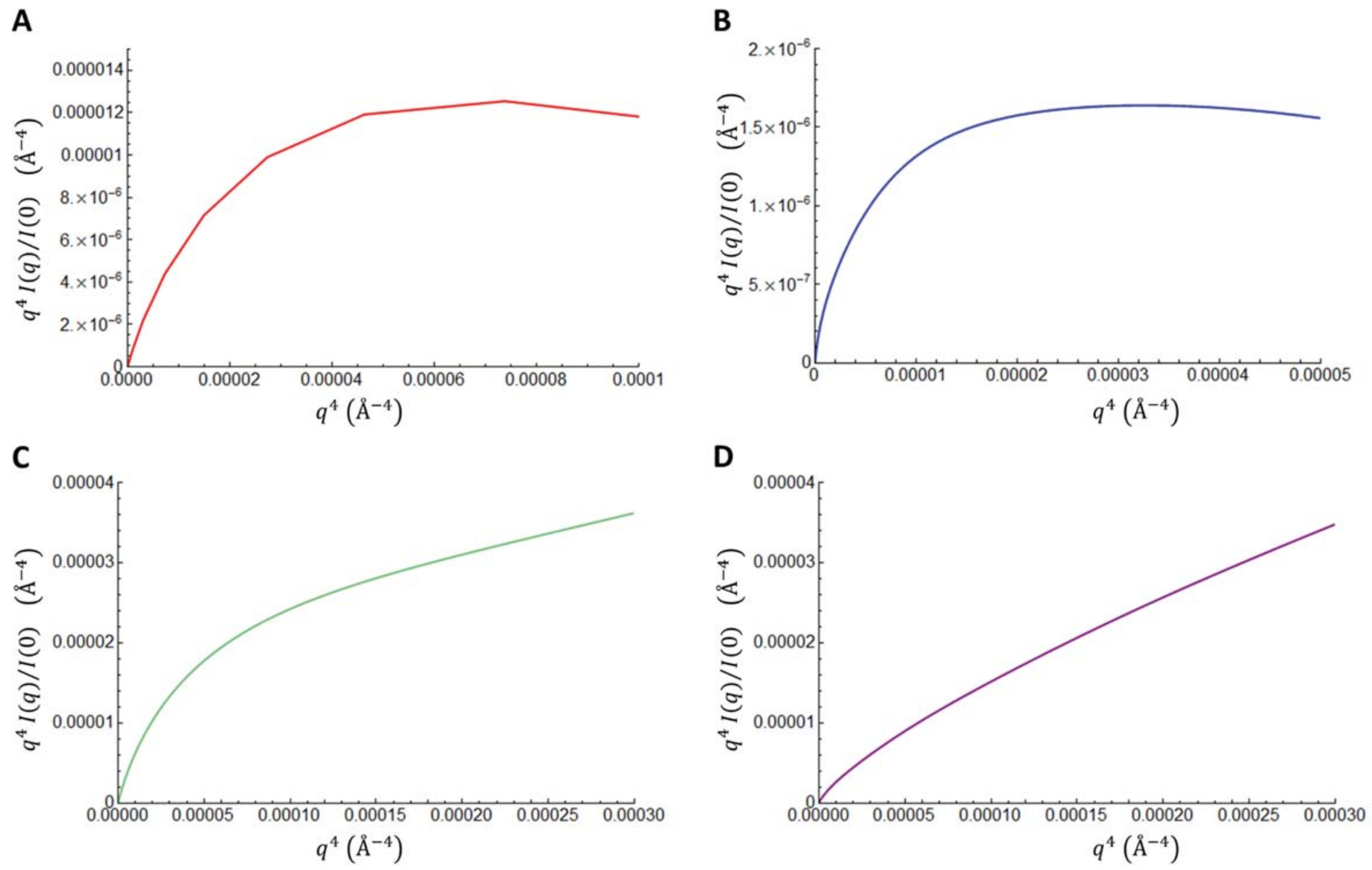

Figure 13: The Porod-Debye plot at low scattering angles for a variety of simple shapes: A) sphere, B) cylinder, C) disk, D) rod.

These simulated results demonstrate that a primary analysis of the scattering data is sufficient to distinguish between globular, nearly spherical, molecules and non-globular molecules. However, in order to infer flexibility, experimental data from other sources is needed. Without corroborating experimental evidence, if the Kratky or Porod-Debye plots indicate that a protein is flexible or spherically asymmetric, it is best to start modeling the "structure" of the protein using the least parameterized modeling approach that can explain the SAXS data - a spherically asymmetric model. Only in cases where this modeling approach fails to adequately describe the experimental data should flexibility be inferred.

2.5.4 A method for describing the statistical conformation of highly flexible proteins

Now that I have discussed a method to determine whether a protein is globular or spherically asymmetric/flexible from the primary SAXS data, without any modeling, I would like to propose a protocol for modeling the structure – the statistical conformation – of highly flexible proteins (Figure 14). This simple protocol uses polymer physics, rigid body, and ensemble modeling approaches in a graduated way that starts from modeling the statistical conformation with a small number of parameters and only increases the "resolution" of the modeling approach when doing so is supported by the SAXS data.

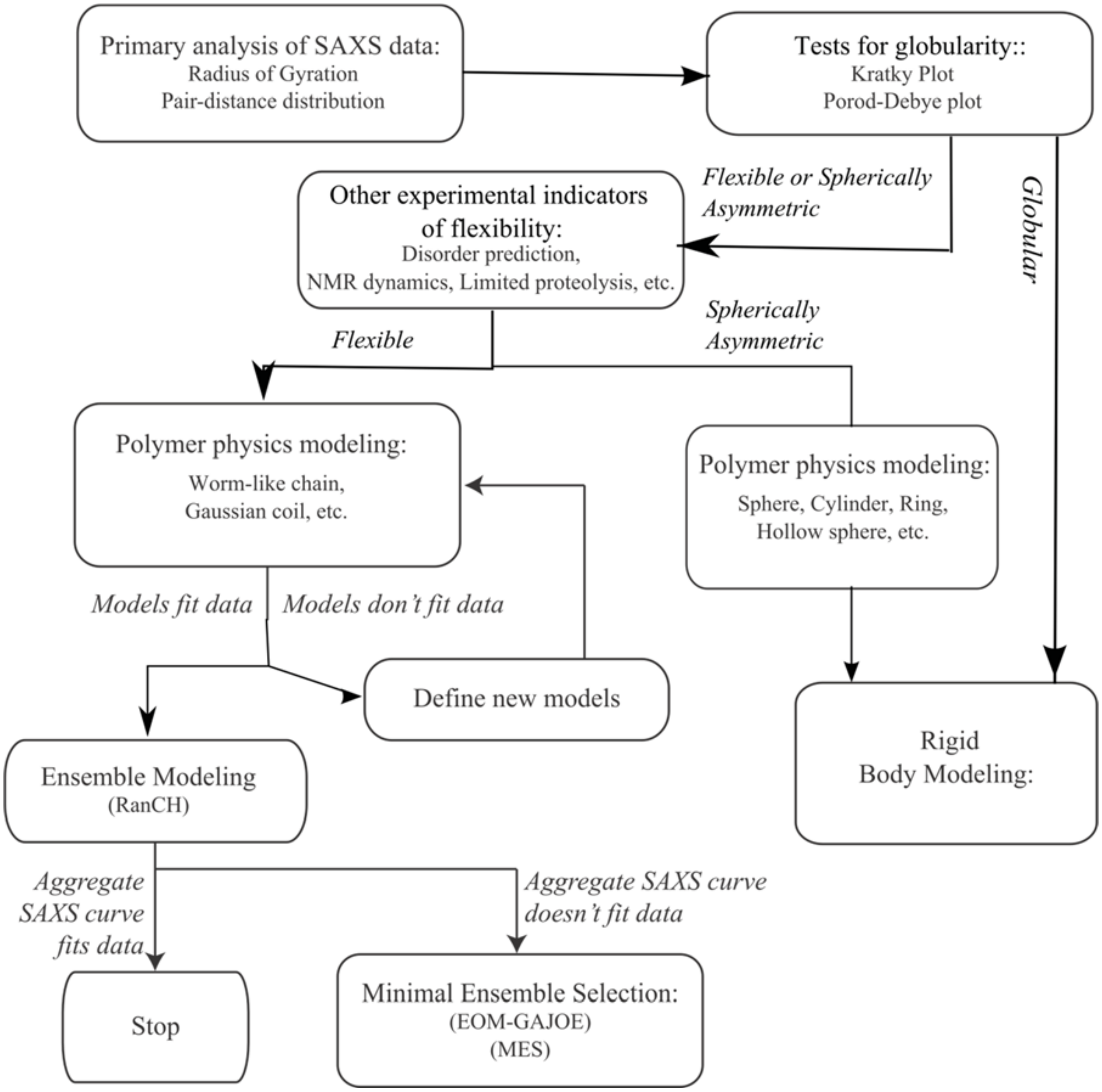

Figure 14: A protocol for modeling highly flexible proteins using SAXS data. When a program is listed, this is the preferred program to use for this step in the protocol.

The first step in this protocol is to determine if the protein is globular or not using the two tests for globularity described above – the dimensionless Kratky plot and the Porod-Debye plot.

1) If the protein is globular, determined by "passing" both the dimensionless Kratky test and the Porod-Debye test, then it is appropriate to model the structure using a rigid body approach.
2) If the SAXS data fails the test for globularity, then two types of modeling should be performed. If there is no additional experimental evidence suggesting that the protein is flexible, polymer physics modeling using rigid simple shapes, or rigid body modeling using atomistic models should be performed. If experimental evidence other than SAXS suggests that the protein is flexible, then polymer physics modeling using flexible models: *e.g.*, worm-like chain, flexible cylinder, Gaussian coil, and pearl necklace should be performed. If no existing polymer model fits the SAXS data, a new model can be defined based on the characteristics of the system. A corresponding scattering function can be derived using an approach similar to the one used in Chapter 4[11].

---

[11] See section 4.5.

3) As a comparison, use an atomistic model and sampling of conformational space to obtain an ensemble representing a coarse sampling of the statistical conformation, unconstrained by the SAXS data. I have developed a Python script that allows one to use RanCH (part of the EOM suite of programs discussed above) to generate an ensemble of models where the ensemble size is dependent on the degrees of freedom of the molecule. This script determines the number of models necessary to produce a converged R_g and D_{max} distribution and generates an aggregate scattering curve for the unconstrained ensemble. If the differences between this calculated scattering curve and the experimental data are random and comparable to noise, then further ensemble analysis is meaningless.

4) If further analysis is indicated, use a program described above to select a minimal set of conformations from the ensemble generated in step 3 whose total predicted scattering curve better fits the observed SAXS data. Then, test this minimal ensemble for statistical significance: Is the R_g distribution determined from the minimal ensemble repeatable using the same starting ensemble? Are the moments of the R_g distribution for each replicate calculation similar? If so, this analysis supports a higher resolution interpretation of the SAXS data, as

represented by the minimal ensemble. If not, the SAXS data does not support such a high resolution depiction of the statistical conformation of the flexible system. In this case, polymer physics provides a depiction that more appropriately reflects the information content of the data.

It is important to note that the generation of the large unconstrained ensemble is an important step in the protocol. The aggregate scattering curve from the large ensemble can be used to confirm the conclusions reached by polymer physics modeling and the global properties of the parent ensemble can provide additional parameters that describe the statistical conformation of the protein. However, one must carefully consider the type of energy functions used to calculate the parent ensemble. I advocate using the simplest method, RanCH or Flexible-Meccano, with the fewest energy functions to generate the ensemble. These are the least parameterized methods for generating the parent ensemble. Using these programs, the aggregate scattering profile is only constrained by hard sphere energy wells. If the aggregate scattering profile does not fit the data, then one can conclude that there are other constraints on the statistical conformation in addition to hard sphere constraints. If another method is used to generate the parent ensemble, it may not be clear precisely what constraints were applied during generation of the ensemble and how these constraints affect the aggregate scattering profile.

2.6 Concluding Remarks

In this chapter I have reviewed the common approaches to modeling the structure of a molecule using experimental SAXS data. Since SAXS is a low information content technique it is very important that the modeling approach used to interpret SAXS data is consistent with information content of the data. The number of parameters in the three most common methods of modeling structures from SAXS data differ, and it is important to start with an approach that has few parameters and only move to more parameterized approaches if the SAXS data warrants it. To facilitate this process, I have outlined a protocol for modeling the statistical conformations of highly flexible proteins. Next I will analyze the statistical conformations of biological macromolecules whose SAXS data have been deposited in public databases (chapter 3), Protein A - a highly flexible multi-domain protein that is a virulence factor in *S. aureus* (chapter 4), and Fibronectin Type III domains 1-2 - a structural protein in the extracellular matrix (chapter 5).

3. Application of the polymer physics analysis to publically available SAXS data.

3.1 Introduction

The protocol presented in Chapter 2 can be widely adopted to analyze the scattering data from potentially flexible macromolecules. This protocol adds an additional domain resolution polymer analysis step to the typical SAXS analysis procedure (Bernado and Svergun, 2012; Putnam et al., 2007). In this chapter, I will test this protocol using publically available SAXS data from an unfolded protein and a minimally flexible multi-domain protein. Using these examples, I will highlight instances where polymer modeling is the appropriate resolution level to describe the SAXS data, and where higher resolution ensemble modeling might be appropriate. I will also discuss what types of information are available from a SAXS analysis of a flexible or elongated molecule, and what types of structural information is not available from a typical SAXS experiment.

3.2 The statistical conformation of an unfolded protein

3.2.1 p15^{PAF} is an intrinsically disordered protein with transient secondary structure elements

p15^{PAF} is a 111 residue long nuclear protein that interacts with proliferating-cell-nuclear-antigen (PCNA). The amino acid sequence suggests that it is an intrinsically disordered protein (IDP) and recent NMR and SAXS studies confirmed that this protein lacks any tertiary structure (De Biasio et al., 2014). De Biasio and colleagues used both SAXS and NMR constraints to generate a statistical conformation of the protein and showed that even though it is an IDP, there are transient non-random secondary structure elements present in the region that binds to PCNA. NMR RDC data was the primary constraint used to generate a representative ensemble of the $p15^{PAF}$ solution conformation, but a radius of gyration (R_g) constraint obtained from SAXS analysis was also used in the modeling.

This study highlights the type of information that one can obtain from conformationally and orientationally averaged SAXS data and what type of information must be determined from other techniques. Although the only SAXS analysis that was performed was a Guinier R_g calculation and a Kratky plot analysis, I will show that by fitting the data to a polymer physics model, one can obtain more detailed information about the statistical conformation.

3.2.2 SAXS analysis of the experimental data

3.2.2.1 Determination of global structural parameters

A global structural parameters obtained from SAXS analysis is the radius of gyration. In the paper, the authors report a Guinier R_g of 28.1 Å ± 0.3 Å obtained to the limit of $qR_g < 1.3$, the limit for globular proteins ((Putnam et al., 2007)). However, for highly flexible molecules the approximated Guinier R_g, which assumes that the molecule is globular, may not appropriately reflect the real R_g for elongated and highly flexible polymers (Jacques et al., 2012; Koch et al., 2003). Using the Debye approximation (Roe, 2000), which assumes that a molecule is highly flexible, I calculated an R_g of 31.3Å ± 0.3Å (Figure 15A). While a 3 Å difference between the Guinier R_g and Debye R_g may not seem significant, since the R_g was the only SAXS constraint used to generate the conformational ensemble, a miscalculation of the R_g almost certainly overly biased the resultant ensemble toward compact conformers.

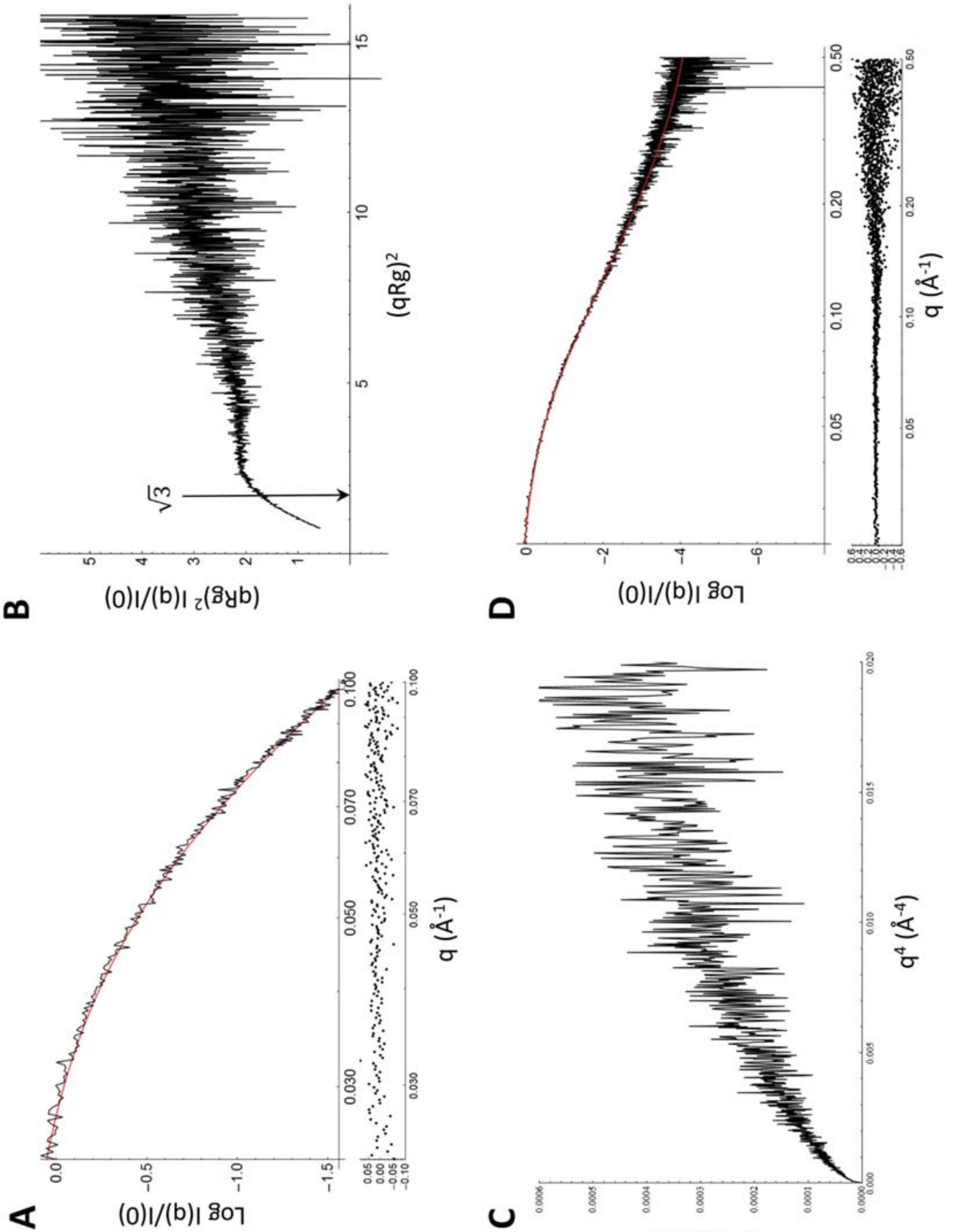
A
Log I(q)/I(0)
0.0
−0.5
−1.0
−1.5
0.05
0.00
−0.05
−0.10
0.030
0.050
0.070
0.100
q (Å^{-1})
B
(qRg)^2 I(q)/I(0)
1
2
3
4
5
$\sqrt{3}$
5
10
15
(qRg)^2
C
q^4 I(q)/I(0)
0.0000
0.0001
0.0002
0.0003
0.0004
0.0005
0.0006
0.000
0.005
0.010
0.015
0.020
q^4 (Å^{-4})
D
Log I(q)/I(0)
0
−2
−4
−6
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
0.05
0.10
0.20
0.50
q (Å^{-1})

Figure 15: p15^{PAF} SAXS data analysis. A) Fit of the Debye function (red) to the low-q region of the data (black). This function is used to determine the R_g of flexible molecules. Residuals are shown below the plot. $\chi^2 = 0.98$. B) The dimensionless Kratky plot does not feature a maxima at $qR_g = \sqrt{3}$, indicating that this protein is non-globular. C) The Porod-Debye plot, which lacks a horizontal plateau, indicates that this protein is non-globular. D) Fit of the excluded volume polymer model (red) to the data (black). Residuals are shown below the plot. $\chi^2 = 0.881$.

3.2.2.2 Polymer modeling

The next step in SAXS analysis is to determine if the protein is globular or non-globular using the dimensionless Kratky plot (Figure 15B) and the Porod-Debye plot (Figure 15C). Both of these plots indicate that the protein may be non-globular. The dimensionless Kratky plot of the data does have a peak at $qR_g = \sqrt{3}$. The Porod-Debye plot does not exhibit a horizontal plateau. The SAXS data was then fit to a minimally parameterized polymer model, in accordance with the protocol outlined in Chapter 2. The best fit polymer model to the data is the excluded volume polymer model, whose approximate scattering function was derived by Hammouda (Hammouda, 1993). The fit of the data to this model is shown in Figure 15D. This polymer model contains two parameters: the R_g of the chain and the Porod exponent. Both the R_g and Porod exponent are a measure of the protein's average size and surface : volume ratio. The Porod exponent is 2.25 ± 0.015, and the R_g is 31.6Å ± 0.17Å. The R_g is in excellent agreement with that calculated from the Debye approximation (31.3Å ± 0.3Å). A Porod exponent of 2.25 suggests that this protein is only slightly more collapsed than a freely-jointed chain (which has a Porod exponent of 2 (Roe, 2000)). The R_g, on the other hand, is slightly larger than that predicted for a 111-residue unfolded protein in good solvent (30.6 Å) (Kohn et al., 2004). However, it must be noted that the calculation for the theoretical R_g of an unfolded protein in good solvent was based on a linear fit of Guinier R_g's for a number of chemically denatured proteins. It may be that the Guinier approximation, which is valid in a lower qR_g region for flexible proteins compared to globular proteins (Jacques et al., 2012), may result in an inaccurate calculated R_g for highly flexible proteins because the approximation implicitly assumes that the protein is globular (Putnam et al., 2007). Both the R_g and Porod exponent indicate that protein:protein interactions in $p15^{PAF}$ are only slightly more preferred than protein:solvent interactions[1]. This minimally parameterized description of the statistical conformation is consistent with the SAXS data and is a more continuous description of the statistical conformation than one that results from enumeration of a discrete ensemble. In summary, an additional structural parameter, the Porod exponent, was determined through polymer modeling and this parameter, along with the R_g, supports the conclusion that $p15^{PAF}$ is a solvated polymer with a slight preference for protein:protein interactions over protein:solvent interactions. Thus, it is reasonable to suggest that transient local

---

[1] A polymer in theta solvent has a Porod exponent of 2, while the Porod exponent of a fully collapsed polymer in poor solvent is between 3 and 4 (Roe, 2000).

structures may be forming in the statistical conformation because intramolecular protein:protein interactions would be consistent with observed SAXS parameters.

De Basio and colleagues also collected ^{1}H-^{15}N RDC data in a stretched poly-acrylamide gel alignment media. They calculated a conformational ensemble consistent with both the NMR and SAXS results. The models in the best-fit ensemble indicated that regions in the center and N-terminus of the protein exhibited transient secondary structure elements. These regions correspond to PCNA and ubiquitin ligase interaction sites. Although additional NMR experiments were used to detect residual secondary structure that was not discernable via SAXS experiments, the description of p15^{PAF} as a polymer with a slight preference for protein:protein interactions is in agreement with De Biasio's conclusions about the "structure" of this protein and the ensemble modeling, but avoids over-fitting the SAXS data (and possibly the NMR data) by enumeration of a discrete ensemble.

3.2.3 Limitations of SAXS

Comparing the information received from SAXS and NMR analysis of this protein highlights what type of information can be obtained from a SAXS experiment, and what type of information is best obtained by other experimental techniques. A typical SAXS analysis of a flexible protein results in parameters that reflect the statistically averaged behavior of a protein, but detailed residue-level information is best obtained from other experimental techniques.

3.3 The statistical conformation of a minimally flexible protein

3.3.1 The domains 1 and 2 of protein tyrosine phosphatase LAR may be minimally flexible.

Bridget Biersmith and colleagues published the crystal structure of the first two domains of the protein tyrosine phosphatase LAR in 2011 (Biersmith et al., 2011). In the crystal structure, both the mouse and drosophila homologs adopt an “unusual horseshoe-like conformation.” This horseshoe-like conformation was present in all members of the asymmetric unit, and buried considerable surface area. In addition, there was a salt bridge present in the model between the two domains. The authors suggested that this salt bridge constrained the conformation of the two domains.

The authors also performed a SAXS analysis to determine if the observed anti-parallel conformation was adopted in solution. The SAXS analysis was complicated by the addition of flexible N- and C-terminal tails, which were not visible in the crystal structure. Ultimately, the authors ended up fitting the SAXS data to a single rigid body atomistic model where the termini adopted a highly extended conformation (Figure 16). The χ^2 goodness of fit statistic of the model to the data was 1.2. The authors concluded that this single model was consistent with both the SAXS data and the crystal structure. However, the sequence of the termini (N: GPGSSRGD, and C: RRVRRVAPRFS) suggests that they do not fully populate an elongated rigid conformation. There may be additional conformations of the termini in solution.

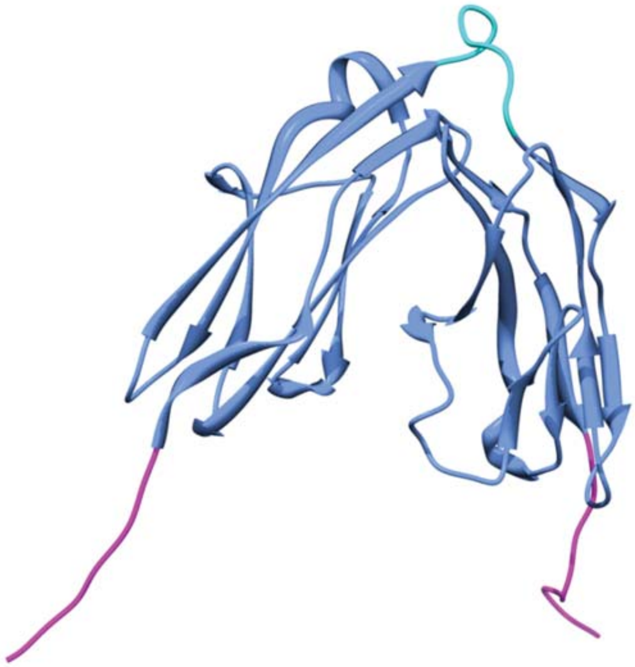

Figure 16: The structure of the first two Ig-like domains of LAR (BioIsis ID: LAR12P). Magenta – the two flexible termini, cyan – a potentially flexible linker.

Though there is strong evidence from the crystal structure that the two domains adopt a predominately horseshoe-like conformation in solution, a SAXS analysis can be useful in detecting sub-populations of elongated or flexible conformers. Additionally, a study of this data will illustrate the limitations of SAXS as a stand-alone technique. I will use polymer modeling and ensemble selection to demonstrate that, in this case, SAXS analysis alone is not able to distinguish between a globular protein with flexible termini and a protein with a flexible inter-domain linker.

3.3.2 SAXS analysis of the experimental data

Both the dimensionless Kratky (Figure 17A) and Porod-Debye (Figure 17B) plots suggest that the protein is not spherical or globular. The dimensionless Kratky plot does not have a peak at $qR_g=\sqrt{3}$, but instead peaks at $qR_g=3$. This plot is similar to the plot for a cylinder or disk, but could also be indicative of a flexible protein. Likewise, the Porod-Debye plot does not have a horizontal asymptote in the low q region, suggesting that the protein is either flexible or spherically asymmetric or both.

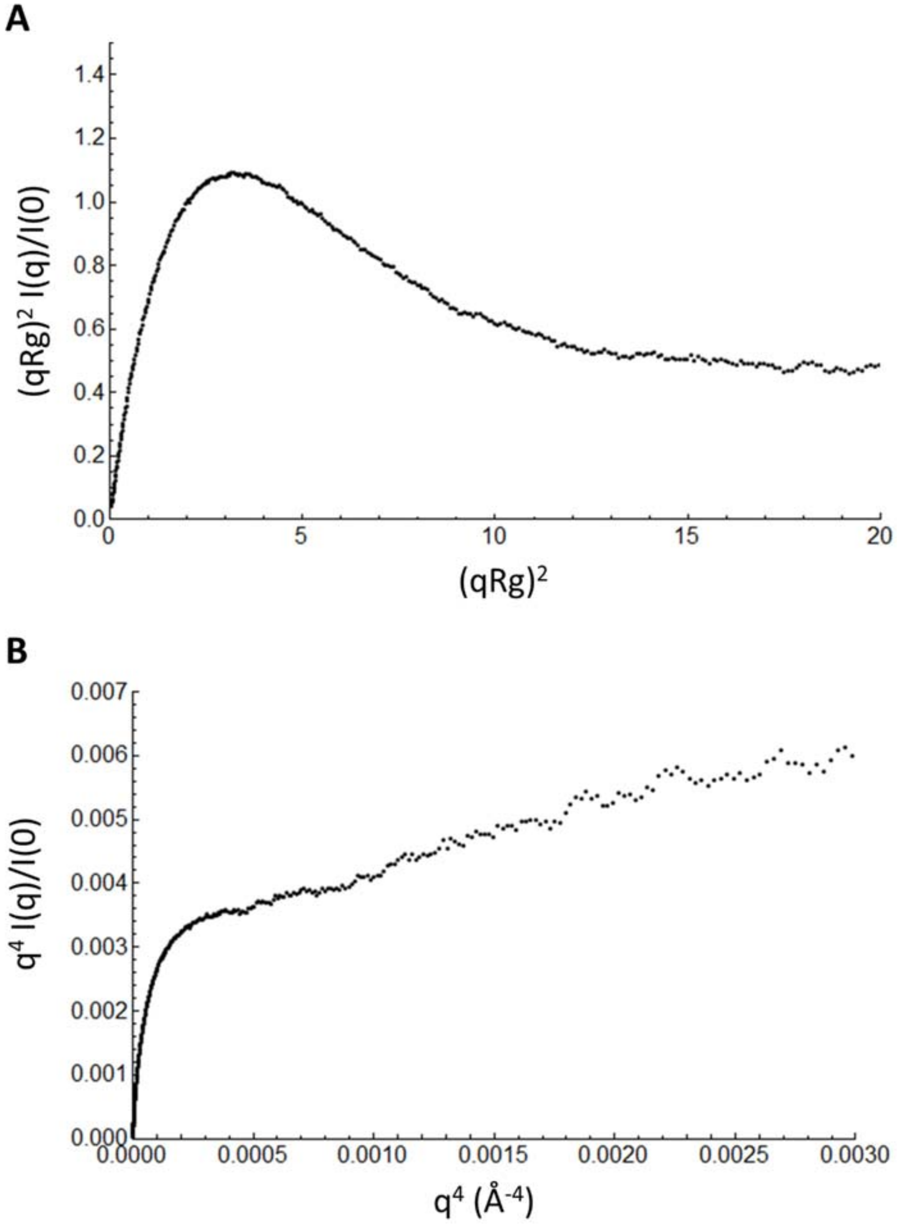

Figure 17: Tests for globularity in LAR SAXS data. A) The dimensionless Kratky plot does not have a peak at qRg = $\sqrt{3}$, indicating that the protein may be non-globular. B) The Porod-Debye plot does not have a horizontal plateau, indicating that the protein may be non-globular.

When I fit the a variety of polymer models to the SAXS data, intriguingly, two different models fit the data with nearly the same fit statistics: the cylinder-ellipse model (Serdyuk et al., 1987) and the flexible cylinder ellipse model (Pedersen, 1996). The cylinder-ellipse polymer model is a static model that contains three parameters: the length of the cylinder-ellipse (19.8Å ± 0.03Å), the major axis radius (31.3Å ± 0.05Å), and the minor axis radius (23.2Å ± 0.05Å). The χ^2 fit statistic to the data was 1.22. The fit of the model to the data and a diagram of this model are presented in Figure 18A.

The flexible cylinder ellipse model fit the data ($\chi^2 = 1.21$) as well as the cylinder ellipse model. Four parameters describe the flexible cylinder ellipse model: the length of the cylinder-ellipse (66Å ± 0.90Å), the Kuhn length (31.6Å ± 0.23Å), the major axis radius (24.9Å ± 0.10Å), and the minor axis radius (9.4Å ± 0.04Å). For flexible chains, the Kuhn length is a measure of the polymer's flexibility (Flory, 1953). The polymer chain is broken down into a series of Kuhn segments of a certain Kuhn length. Each segment is considered to be freely-jointed with the neighboring segments. The LAR flexible cylinder ellipse model is broken down into two Kuhn segments (diagrammed in Figure 18B, bottom). The fit of the flexible cylinder ellipse model to the data is presented in Figure 18B, top.

Notably, the χ^2 statistics of the polymer models (1.22 and 1.21) are nearly identical to the fit-statistic from the single rigid body model (1.2). This result highlights the lack of information content in the SAXS data. Three different models can describe the SAXS data equally well. However, a χ^2 of 1.2 is rather large, indicating that there are features in the SAXS data not accounted for in any of the models. Ensemble modeling may result in a better description of the statistical conformation than either the atomistic rigid body model or the polymer models.

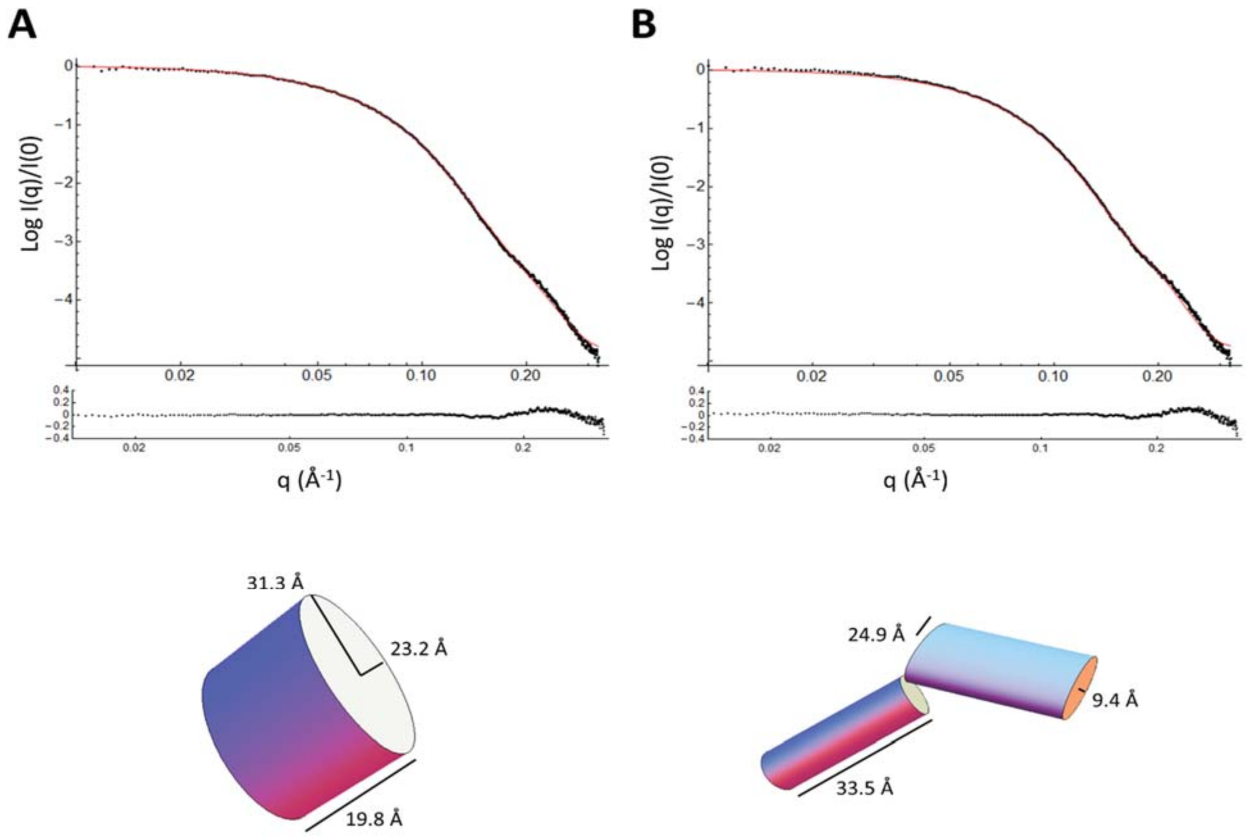

Figure 18: Fit of polymer models to the LAR SAXS data. A) Top: fit of the cylinder ellipse model (red) to the SAXS data (black). $\chi^2 = 1.22$. Residuals are below the plot. Bottom: diagram of the cylinder ellipse model with best fit parameters indicated. B) Top: fit of the flexible cylinder ellipse model (red) to the SAXS data (black). $\chi^2 = 1.21$. Residuals are below the plot. Bottom: diagram of the flexible cylinder ellipse model with best fit parameters indicated.

3.3.3 Ensemble modeling of LAR domains 1 and 2.

To test whether LAR is best described as a rigid protein with flexible linkers or as a flexible multi-domain protein with flexible linkers, two different conformational ensembles were constructed using RanCH (Bernado et al., 2007) in random coil mode. In the first ensemble (the *flexible* ensemble) the N- and C- termini and the seven residue inter-domain linker were allowed to adopt random conformations and the domain orientation of the two rigid domains was unconstrained. In the second ensemble (the *fixed* ensemble) the N- and C-termini were allowed to adopt random conformations but the seven residue inter-domain linker and the domain orientations were fixed in the orientation that was observed in the crystal structure.

The aggregate SAXS curves from each 10,000 member ensemble are presented below in Figure 19. The fit of the aggregate SAXS profile from the flexible ensemble to the data is poor ($\chi^2 = 5.33$) (Figure 19A). This poor fit is expected since the best-fit polymer model was a semi-flexible polymer instead of a highly flexible one. Because only hard-sphere energy functions are used to constrain each conformer in the RanCH-generated ensemble, ensemble will have the same global characteristics as a highly flexible excluded-volume coil.

The fit of the aggregate fixed ensemble to the SAXS data is better than the fit of the flexible ensemble ($\chi^2 = 1.30$) (Figure 19B). However, particularly in the high q region, where one would expect to observe a good fit of an atomistic model to the SAXS data, both low resolution polymer model profiles fit the SAXS data better than the aggregate profiles from the large atomistic ensembles. The additional step of enumeration of a minimal atomistic ensemble is indicated.

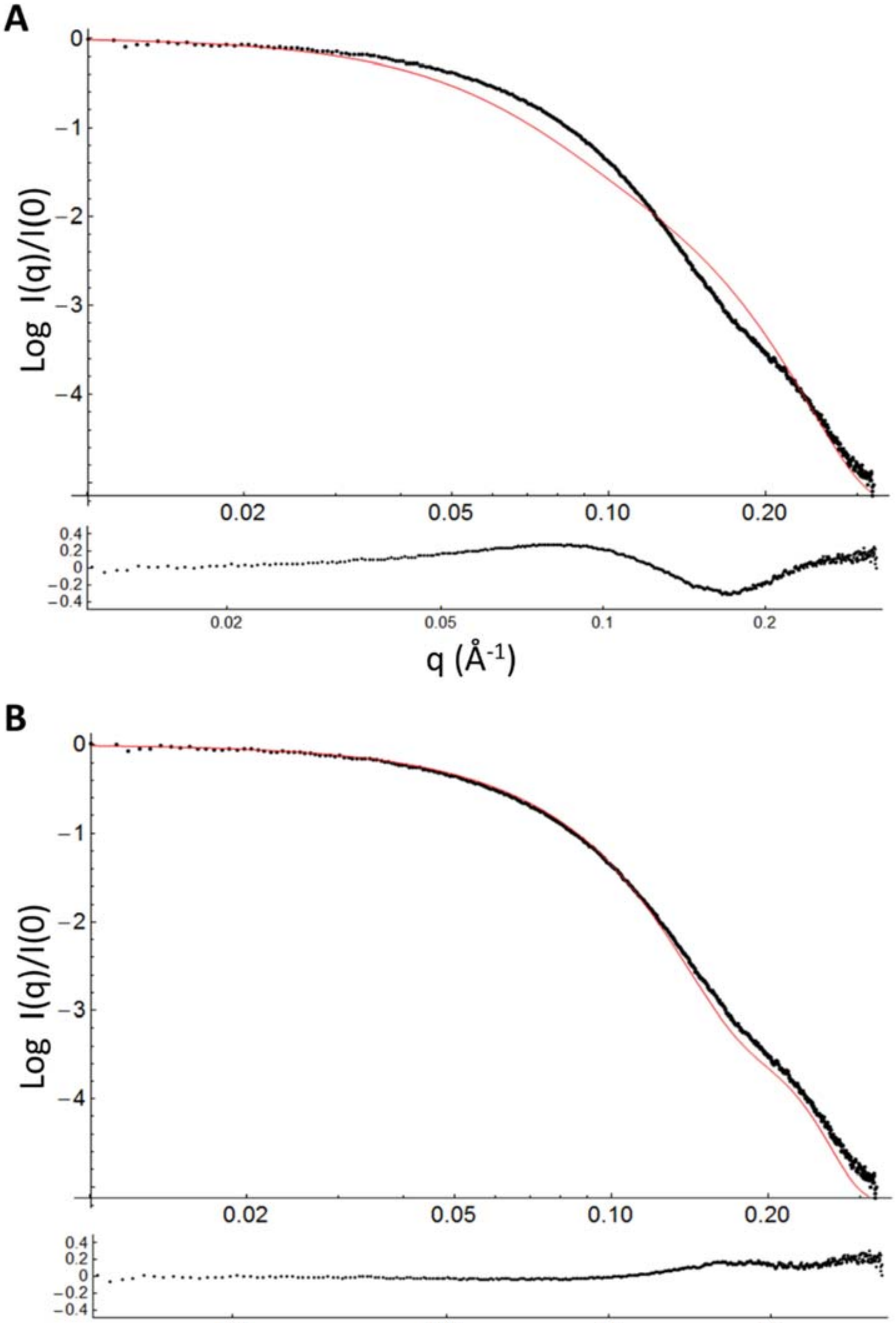
A
Log I(q)/I(0)
0
−1
−2
−3
−4
0.02
0.05
0.10
0.20
0.4
0.2
0
−0.2
−0.4
0.02
0.05
0.1
0.2
q (Å-1)
B
Log I(q)/I(0)
0
−1
−2
−3
−4
0.02
0.05
0.10
0.20
0.4
0.2
0
−0.2
−0.4
0.02
0.05
0.1
0.2
q (Å-1)

Figure 19: Fit of the aggregate theoretical SAXS profiles from 10,000 member ensembles to the LAR SAXS data. A) The aggregate SAXS profile of the flexible ensemble (red) and the experimental data (black), $\chi^2 = 5.33$. Residuals are shown below the plot. B) The aggregate SAXS profile of the fixed ensemble (red) and the experimental data (black), $\chi^2 = 1.30$. Residuals are shown below the plot.

GAJOE, part of the EOM suite of programs (Bernado et al., 2007), was used to calculate minimal ensembles that best fit the LAR SAXS data from both the fixed and flexible parent ensembles. Three minimal ensembles were generated for each parent ensemble in order to assess uniqueness and robustness of the final minimal ensembles. The results are presented in Figure 20, below.

The theoretical SAXS profiles from each minimal ensemble chosen from the flexible parent ensemble were identical, and the residuals of each fit were stochastically distributed about the mean in the high q region of the scattering profile. (Figure 20A, top. Residuals are shown for each minimal ensemble below the plot in red, green, and blue.) The theoretical SAXS profile of the minimal ensembles were excellent fits to the experimental data ($\chi^2 = 1.07$). The R_g and maximum distance (D_{max}) distributions of each minimal ensemble were statistically identical, but were distinct from the R_g and D_{max} distribution of the parent ensemble (Figure 20A, middle and bottom). In all cases, the best fit minimal ensembles were composed primarily of compact models, however, in each ensemble there was a small population of more elongated conformations, which is reflected in the tails of the R_g and D_{max} distributions in Figure 20A. This result suggests that the solution conformation of LAR may not fully populate the conformational space of the “horseshoe.” Instead, there might be a small population of more elongated open conformations that exists in equilibrium with the horseshoe conformation. An interesting follow-up experiment would be to test the ionic-strength dependence of the R_g and D_{max}. If the salt bridge observed in the crystal structure is constraining the statistical conformation, then a change in ionic strength of the solution should change the observed R_g and D_{max} distributions of the minimal ensembles.

The fit of the theoretical SAXS profile from the minimal ensembles generated from the fixed parent ensemble was only marginally poorer than those from the flexible ensemble ($\chi^2 = 1.09$). As expected, the R_g and D_{max} distributions of both the fixed parent and minimal ensembles were much smaller and had more kurtosis than the flexible ensembles (Figure 20B). Recall that the fixed parent ensemble was generated by fixing the domain orientation so that only the termini were flexible. By allowing the termini to adopt more than one conformation, the fit of the theoretical SAXS profile to the data was improved compared to the original rigid body model proposed by Biersmith and colleagues (fit of the original model to the data: $\chi^2 = 1.2$, fit of the ensemble models to the data: $\chi^2 = 1.09$). The residuals of the ensemble fits are non-stochastic in the high q region (Figure 20B, top). Since the flexible ensemble fits the experimental data better than the fixed ensemble, there is at least some flexibility in the inter-domain linker.

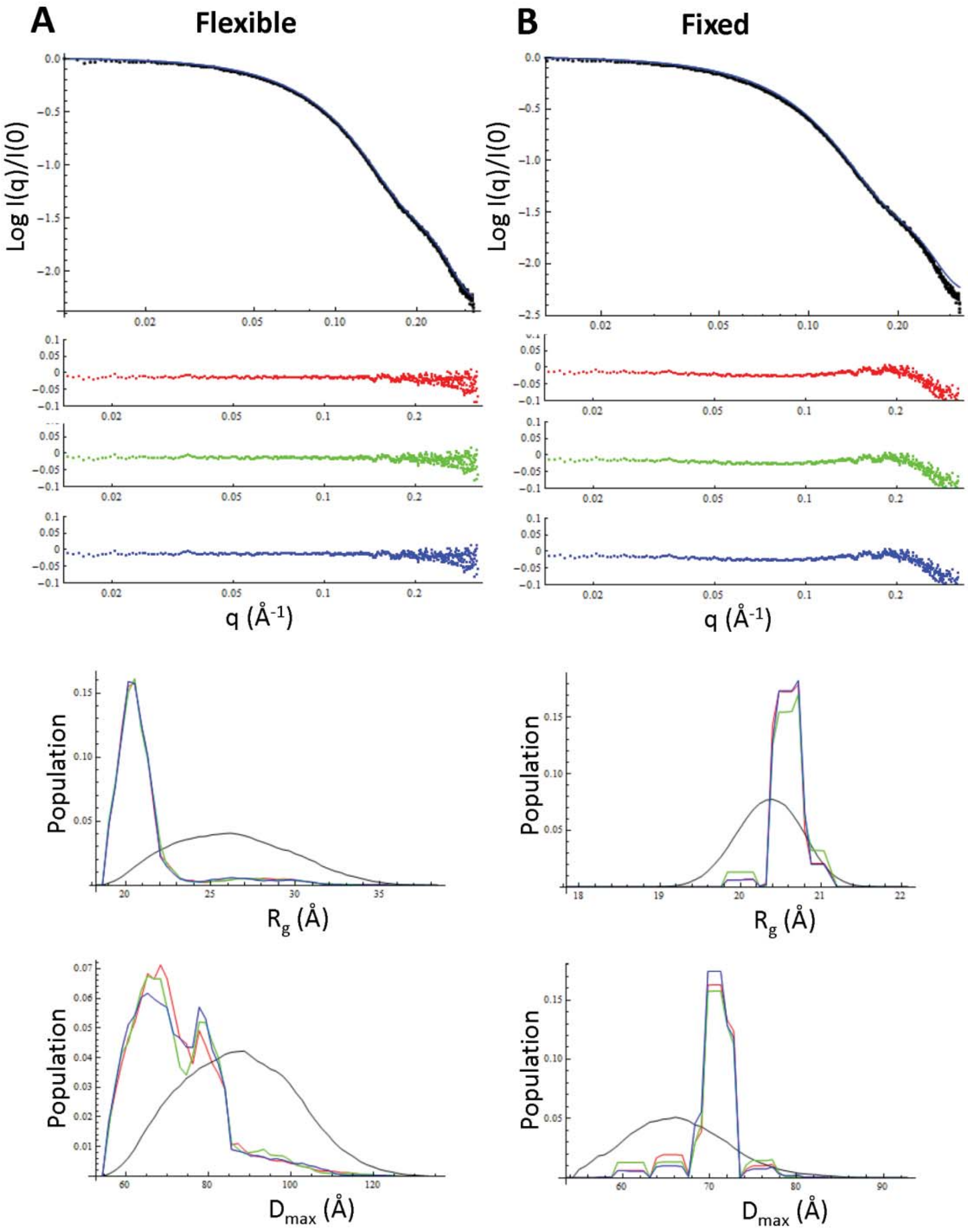

Figure 20: The GAJOE results for the three minimal ensembles calculated from the flexible (A) and fixed (B) parent ensembles. Top: Fit of the minimal ensemble SAXS curve to the data (black), flexible χ^2 = 1.07, fixed χ^2 = 1.09. Residuals are shown below the plot for each minimal ensemble generated (red, green, blue). Middle: The R_g distribution of each minimal ensemble (red, green, blue) and the parent

distributions (black). Bottom: The D_{max} distribution of each minimal ensemble (red, green, blue) and the parent distributions (black).

3.3.4 The flexibility of LAR cannot be determined solely by SAXS analysis

From the results presented above, it is reasonable to suggest that there is some flexibility in the LAR statistical conformation. The goodness-of-fit statistics for all the models are presented in Table 1. The models that can be immediately ruled out by considering the SAXS data alone are the single rigid body model ($\chi^2 = 1.2$) and the aggregate flexible parent ensemble where the domain orientations are not constrained ($\chi^2 = 5.33$). The fit of the flexible models to the SAXS data suggests that there is at least some flexibility in the linker, but other experiments are needed to characterize the extent of this flexibility.

Table 2: The X^2 statistic for the fit of each model in section 3.3 to the LAR SAXS data.

Model	Reduced X^2 Statistic
Original rigid body model	1.20
Cylinder-ellipse polymer model	1.22
Flexible cylinder-ellipse polymer model	1.21
Aggregate Flexible parent ensemble	5.33
Aggregate Fixed parent ensemble	1.30
Aggregate minimal Flexible ensemble	1.07
Aggregate minimal Fixed ensemble	1.09

The complete SAXS analysis of the LAR dataset highlights some important limitations in determining models using SAXS data. The ambiguity in the polymer modeling results emphasizes that just because a model is consistent with SAXS data, it is not necessarily unique.

3.4 Concluding Remarks

The low-information content of SAXS data and the importance of appropriate modeling methods are highlighted in the above example. In the case of the intrinsically disordered protein, a polymer analysis was able to determine that the statistical behavior of the protein in solution was consistent with that of a flexible polymer with a slight preference for protein:protein interactions over protein:solvent interactions. Additional NMR analysis was able to provide a higher resolution description of the statistical conformation than SAXS analysis alone. Both the SAXS analysis and NMR analysis were used conjointly to determine a minimal ensemble consistent with the data. In the case of the minimally flexible protein, LAR, SAXS analysis demonstrated that the statistical conformation was flexible. More experimental studies are needed to determine the extent of flexibility in the linker.

These examples illustrate the way in which small-angle x-ray scattering can aid in describing the statistical conformation of biological macromolecules. Very rarely can SAXS alone be used to fully describe the statistical conformation. Instead, its' utility is best used in coordination with other methods. That being said, the following two chapters will illustrate that functionally insightful descriptions of the statistical conformation can be obtained from SAXS, and that other experimental techniques verify the conclusions drawn from a SAXS analysis.

3.5 Methods

3.5.1 Analysis of p15^{PAF} SAXS data

Calculation of the Debye R_g was performed in SASView using the Debye model of a polymer chain (Roe, 2000). The q-region used for this analysis was 0.024 – 0.10 Å^{-1}, consistent with the reported “Debye region” for flexible polymers (Putnam et al., 2007). The fit of the SAXS data to the excluded volume polymer model was also performed in SASView. This program uses non-linear lest squares fitting to fit the data to the model and minimizes the reduced χ^2 statistic (Svergun et al., 1995).

3.5.2 Analysis of LAR SAXS data

Comparison of polymer models and fit of the models to the experimental SAXS data was performed in SASView. Calculation of the χ^2 statistic for all models was performed in Mathematica 9 using the reduced χ^2 statistic reported by Svergun in 1995 (Svergun et al., 1995).

Generation of the parent fixed and flexible ensembles were performed in RanCH (part of the EOM suite of programs). For the flexible ensemble, the atomistic model from the BioIsis database (BioIsis ID: LAR12P) was used as a starting structure. The two IgG domains were entered as separate models, and the conformations of the 7-residue linker and the termini were randomized. RanCH generated 10,000 structures in random coil mode. For the fixed ensemble LAR12P was used as a starting model. The atomistic model was trimmed to eliminate the N- and C-termini and was entered as a single model. RanCH randomized the termini and generated 10,000 models in random coil mode. After the models were generated, CRYSOL was used to generate theoretical the scattering profile for each model. 200 q values were used for each profile, and the scattering profiles were approximated using 15 spherical harmonics. After the parent ensemble was generated, GAJOE (part of EOM) selected minimal ensembles. 1000 generations were performed for each GAJOE cycle, and the best-fit ensembles from 100 cycles were compared, and the ensemble with the lowest χ^2 statistic was reported. For each parent ensemble, this GAJOE procedure was performed three times.

4. The statistical conformation of a highly flexible protein: *S. aureus* protein A

4.1 Introduction

The preceding chapter illustrated, through the analysis of publically available SAXS data, the type of information one can reasonably obtain from a SAXS analysis. In some cases, the SAXS data is unambiguous – one model clearly fits the experimental data better than other models. In more complex systems, due to the low information content of the data, it may be difficult to distinguish between two or more very different models that describe the data equally well. In this chapter, I have applied this knowledge and the analysis protocol in Chapter 2 to the study of an experimental system: *Staphylococcus aureus* protein A.

Staphylococcus protein A (SpA) functions as a crucial *S. aureus* virulence factor through a wide array of intermolecular interactions (Palmqvist et al., 2002). It has been shown to bind to the Fc fragment of antibodies to inhibit host immune response (Deisenhofer, 1981; Moks et al., 1986). It can activate TNFα receptors (Gomez et al., 2004), leading to the inflammatory response, sepsis and death of the host. It binds to von Willebrand factor (Hartleib et al., 2000), allowing *S. aureus* to adhere to platelets and withstand shear stress. In addition, it binds to C1qR inhibiting complement pathway activation and the host immune response (Nguyen et al., 2000). SpA also plays a role in biofilm formation (Merino et al., 2009), although the exact mechanism is unknown. As in other systems, this diverse range of SpA functions may be associated with its structural flexibility. A description of SpA's structural flexibility will facilitate a better understanding of the role this property plays in the protein's diverse functions. Within the definition of flexibility, there are two extremes of conformational flexibility to consider: intra-domain local motions, consisting of side-chain flexibility and movement of secondary structure elements, and inter-domain global motions, ie: the movement of one domain relative to the others. This study focuses on a description of the inter-domain flexibility in SpA.

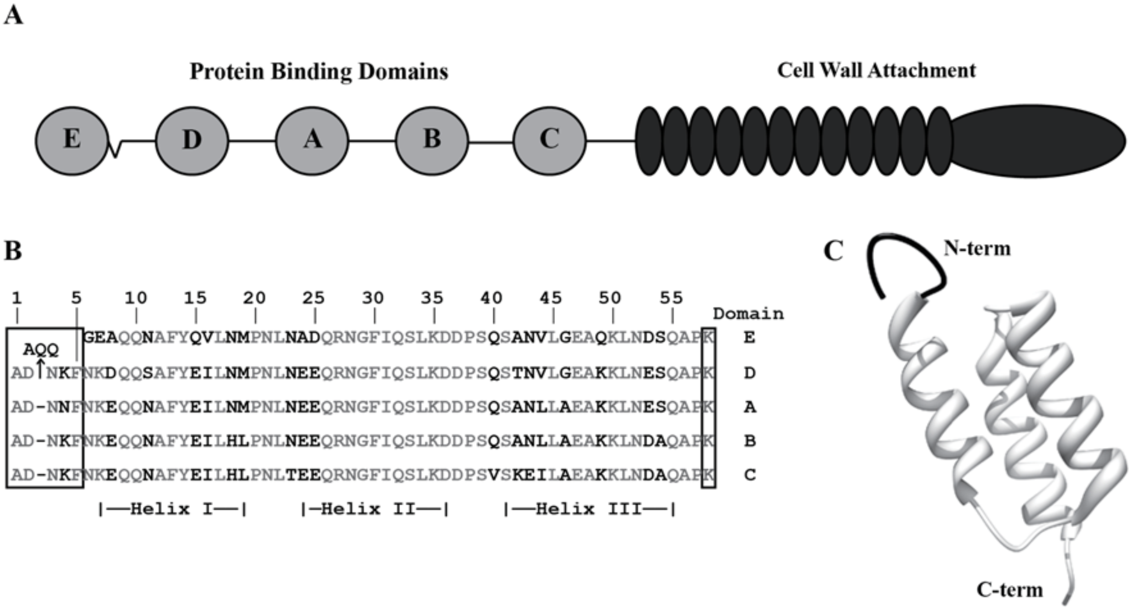
A
Protein Binding Domains
Cell Wall Attachment
E
D
A
B
C
B
1 5 10 15 20 25 30 35 40 45 50 55
Domain
AQQ
GEAQQNAFYQVLNMPNLNADQRNGFIQSLKDDPSQSANVLGEAQKLNDSQAPK E
AD NKFNKDQQSAFYEILNMPNLNEEQRNGFIQSLKDDPSQSTNVLGEAKKLNESQAPK D
AD-NNFNKEQQNAFYEILNMPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNESQAPK A
AD-NKFNKEQQNAFYEILHLPNLNEEQRNGFIQSLKDDPSQSANLLAEAKKLNDAQAPK B
AD-NKFNKEQQNAFYEILHLPNLTEEQRNGFIQSLKDDPSVSKEILAEAKKLNDAQAPK C
|—Helix I—|
|—Helix II—|
|—Helix III—|
C
N-term
C-term

Figure 21: ***Staphylococcus aureus*** **protein A (SpA). A) A schematic of SpA shows its' two major regions: an N-terminal protein binding region (SpA-N) and a C-terminal domain involved in cell wall attachment. There is a conserved linker (black) between each of the 5 protein binding domains, E, D, A, B, and C (light grey). B) Sequence alignment of the five nearly-identical protein binding domains. Sequence identical in all domains is shown in grey. The linker region is boxed in black. There is**

a 3 amino acid insertion between domains E and D. C) The structure of Z-BdpA, a B domain homolog. pdb: 1Q2N. Each SpA-N domain consists of a three helix bundle and flexible N- and C- termini (shown in black).

SpA is a multi-domain protein consisting of an N-terminal domain containing a signal sequence and five protein binding domains (Lofdahl et al., 1983; Moks et al., 1986), and a C-terminal region used to target the protein to the cell surface via an LPXTG motif (Schneewind et al., 1995) (Figure 21A). The N-terminal half of this protein (SpA-N) interacts with the host-cell proteins, mediating the immune response. The five protein binding domains in the N-terminal half are the functional portion of the protein and have a high degree of sequence identity (Figure 21B). The structure of D domain has been determined in complex with the Fab fragment of a human IgM antibody (Graille et al., 2000). The structures of E and B domains have been determined by NMR spectroscopy (Starovasnik et al., 1996; Zheng et al., 2004). The 58-residue domains consist of three almost-parallel alpha-helices with N- and C-terminal flexible residues (Figure 21C). The Oas lab at Duke University has previously determined that the five domains fold independently of each other; there is no thermodynamic coupling between the domains. An NMR dynamics map indicates that there is a six-residue flexible linker between each domain.

What is unknown is how the individual domains are structurally related to each other and how the presence of repeated domains structurally constrains the protein. It is these structural constraints that constrain the flexibility of the statistical conformation and affect the thermodynamics of the statistical conformation. So, by determining the statistical conformation we can gain important insights into how it contributes to the function of SpA-N.

In this study, I used small angle x-ray scattering (SAXS) to determine and describe the statistical conformation of SpA-N. In order to study the statistical conformation of SpA-N and determine the structural relationship between individual domains William Franch (an Oas lab member) constructed a series of proteins consisting of 1, 2, 3, 4, or 5 repeats of the B-domain (BdpA). This simplification of SpA-N into five identical domains allows me to use a polymer physics approach to describe the statistical conformation. A polymer physics approach is one that seeks to describe the statistical properties of a polymer using simple mathematical models. From these simple polymer models, I can describe the "structure" of the thermodynamic state of the ensemble – the statistical conformation.

4.2 The SAXS analysis shows that SpA-N can be described as a polymer

SAXS data were obtained from five B-domain protein fragments (BdpA, 2-BdpA, 3-BdpA, 4-BdpA, 5-BdpA) and the N-terminal region of SpA (SpA-N). This allowed me to use a combinatorial approach to study the statistical conformation of SpA-N and derive analytical models to describe that conformational space. Five data sets were collected for each protein fragment at a concentration range of 5 - 0.5 mg/ml. Data sets were screened for concentration dependent effects in the low-q region, and the data sets for each protein fragment that were free of concentration dependent effects and had the highest signal : noise ratio were selected for further analysis.

4.2.1 The radius of gyration of SpA-N and N-BdpA can be fit to an excluded volume polymer model

A Guinier analysis of the scattering data allows for direct estimation of the radius of gyration (R_g) of each protein construct (Koch et al., 2003). The Guinier plot is an algebraic transformation (ln(I) vs. q^2) of the data that produces a linear q^2 dependence in the "Guinier region" found at very small scattering angles ($q < 0.05$ $Å^{-1}$). The slope of the data is directly proportional to the radius of gyration of the overall protein chain. The q-range of the Guinier region is dependent on R_g and the globularity of the molecule (Hjelm, 1985). We determined the Guinier region and R_g of 1-BdpA, which is globular, to the limit of $q^*R_g < 1.3$. The Guinier regions and R_g of (2-5)BdpA, and SpA-N were determined to a $q^*R_g < 1.0$, the limit of the Guinier region for elongated or flexible macromolecules (Hjelm, 1985; Jacques et al., 2012) (Figure 22).

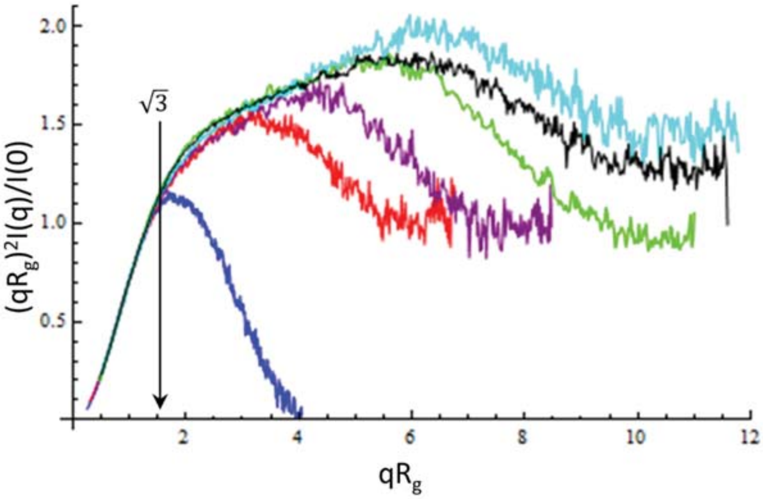

Figure 22: The dimensionless Kratky plot indicates that (2-5)-BdpA and SpA-N may not be globular. 1-BdpA converges to 0 at $q^*R_g = 4$, indicating that it is a globular protein. The Kratky plots of (2-5)-BdpA and SpA-N do not return to zero, indicating that these molecules are flexible or spherically assymetric. Shown: A dimensionless Kratky plot of N-BdpA. Blue: 1-BdpA, Red: 2-BdpA, Purple: 3-BdpA, Green: 4-BdpA, Cyan: 5-BdpA, Black: SpA-N.

Excellent linear correlations within the Guinier regions are observed for 2-BdpA, 3-BdpA, 4-BdpA, 5-BdpA, and SpA-N data (Figure 23), indicating that each sample was free of self-association or interparticle interference, which might otherwise bias derived models. Interparticle interference is observed in the very low-q region of 1-BdpA, so only the data from $0.0005 < q^2 < 0.01$ was used to estimate R_g. The Guinier plots for the

N-BdpA protein fragments and SpA-N show a systematic increase in R_g as domains are added (Figure 23). However, the R_g is not a linear function of the number of domains, indicating that 5-BdpA and, by extension, SpA-N are not elongated rigid rods, but rather can be described by a polymer model (see below). The SpA-N R_g is 6.8% smaller than that of 5-BdpA, indicating that SpA-N is more compact than 5-BdpA, even though the two molecules have nearly identical molecular weights (32571 Da for SpA-N vs. 33186 Da for 5-BdpA, 1.8 % different). This difference may be the result of inter-domain interactions that are more favorable in SpA-N or more unfavorable in 5-BdpA, or both.

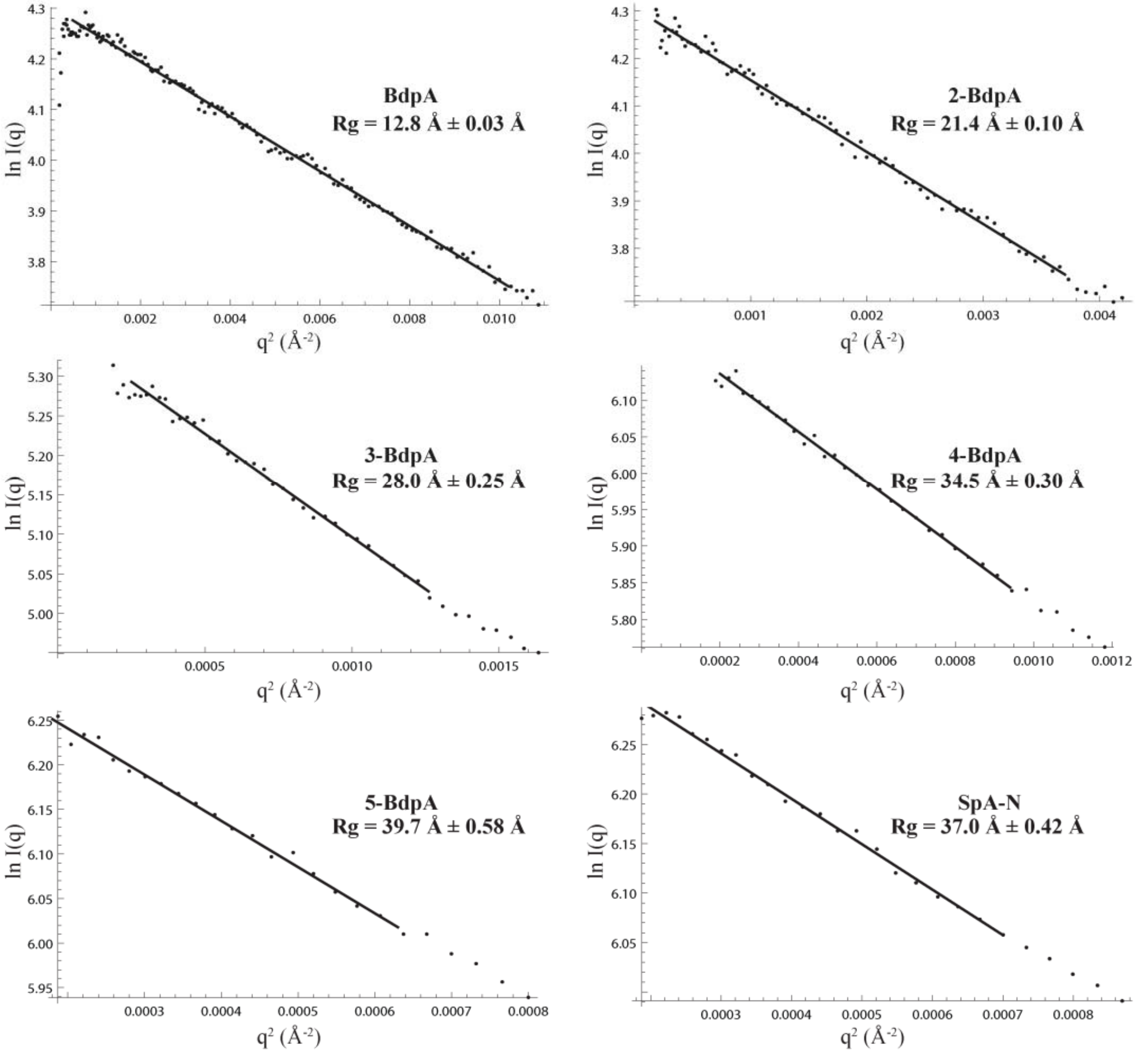

Figure 23: Guinier analysis of N-BdpA and SpA-N protein fragments. The Guinier plots for all constructs show excellent linear correlations in the low-q regions (solid lines). The radius of gyration is not a linear function of monomer number. This indicates that N-BdpA and SpA-N do not explore a conformational space consistent with that of a rigid rod, but rather can be fit to a polymer model. Protein concentrations for each dataset: BdpA - 1 mg/ml, 2-Bdpa - 1 mg/ml, 3-BdpA - 5mg/ml, 4-BdpA - 5 mg/ml, 5-BdpA - 5 mg/ml, SpA-N - 5mg/ml.

The systematic increase in R_g among the N-BdpA protein fragments suggests that the relationship between monomer number and R_g can be fit to a polymer model. A simple model that describes the stiffness and conformational space of a polymer is the swollen Gaussian coil. For this model, the R_g is (Hammouda, 1993):

$$R_g = l_p\sqrt{\frac{N^{2v}}{(2v+1)(2v+2)}} \qquad 4.1$$

where l_p is the persistence length of the polymer and ν is the Flory coefficient. The persistence length gives the length scale of polymer flexibility, which is reflected in the Flory coefficient. A freely-jointed chain has a persistence length on the same order of magnitude as the bond length and $v = 0.5$ (Flory, 1953). When the persistence length is on the same order as a bond length and $0.5 < v < 1$, then the polymer is said to be semi-flexible (Rubinstein and Colby, 2003). A rigid rod has a persistence length of ∞ and $v = 1.0$. It is important to note that this model allows long polymers to intersect with themselves but the avoidance of short-range monomer intersections is accounted for in the parameters l_p and v, which reflect chain stiffness, i.e., short-range excluded volume effects.

A non-linear least-squares fit of the N-BdpA Guinier R_g vs. monomer number to Equation 4.1 gives a persistence length of 37.5 Å (95% confidence intervals: 36.3-38.8 Å) and a Flory coefficient of 0.68 (95% confidence intervals: 0.64-0.72) (Figure 24). The persistence length is comparable to the length of a single BdpA (29.6 Å) and three flexible residues (10.2 Å, assuming 3.4 Å is the average distance between Cα atoms in an unstructured polypeptide). The best-fit Flory coefficient is larger than the coefficient for a fully-swollen Gaussian coil (0.6) (Flory, 1953)but smaller than that of a rigid rod (1.0), indicating that 5-BdpA, and by extension SpA-N, behaves as a semi-flexible excluded volume biopolymer, not as a rigid rod or an ideal Gaussian coil.

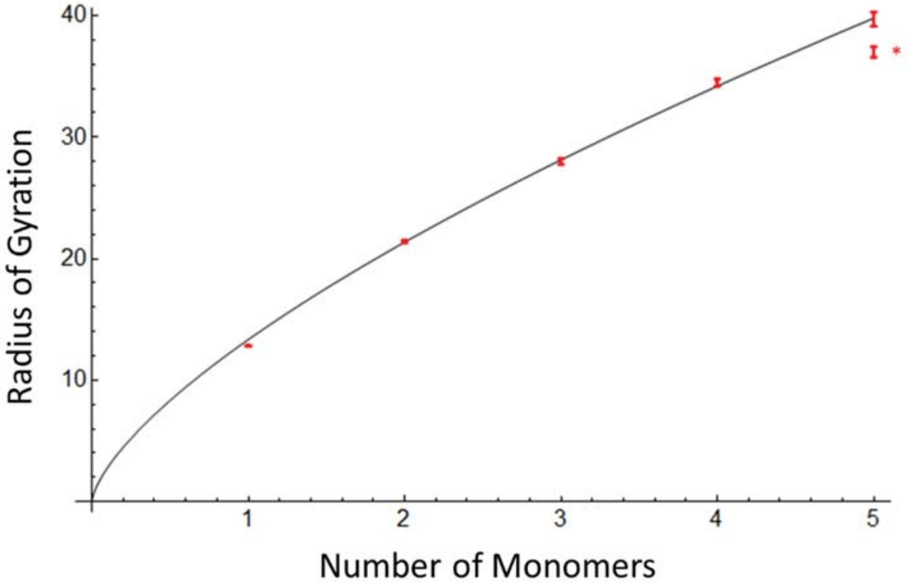

Figure 24: Non-linear least squares fit of the radius of gyration of N-BdpA (red) to equation 4.1 (black). The radius of gyration does not increase linearly with increasing monomer number. The excellent agreement of the data with equation 4.1 suggests that N-BdpA behaves as a flexible polymer. * indicates the radius of gyration for SpA-N.

4.2.2 The statistical conformation of 5-BdpA can be described by the excluded volume pearl necklace polymer model.

The radius of gyration is a model-independent measure of the polymer's overall size. It is calculated from the very low-q region of the scattering curve. The higher-q region of the scattering curve can be used to determine the overall shape and behavior of the polymer by comparing scattering curves of simple shape models to the scattering data. For example, the SAXS scattering curve of simple shapes, such as a sphere, cylinder, worm-like chain, and Gaussian coil, can be used to discriminate between polymer models. Even the scattering curves of a subset of simple shapes can be used to distinguish between polymer models.

One set of polymer models is the Gaussian coil models, where the Flory coefficient and statistical segment length describe the flexibility of the polymer, and the end-to-end distance distribution of the polymer is a Gaussian distribution. The swollen Gaussian coil model describes the monomers as points whose spatial arrangement is that of a random coil, and excluded volume interactions are taken into account by an increase in the Flory coefficient where $v > 0.5$ (the Flory coefficient for a random flight polymer). In the scattering function for this model (Hammouda, 1993) there is no term to account for the volume of the monomers except for the statistical segment length and the Flory coefficient.

A subset of Gaussian coil polymer models is the pearl necklace model (PNM). This model has been used to describe polyelectrolytes in various solvents (Dobrynin et al., 1996) and the statistical conformation of long repeat proteins like fibronectin (Pelta et al., 2000). In the pearl necklace model the monomers are represented as spheres separated by a linker. Previous studies have used various implementations of this model to derive scattering functions for fitting scattering data. These variations include different equations for the relative positions of the "pearl" monomers (structure factor) and the contribution of the linker to the scattering. In contrast to the swollen Gaussian coil model, pearl necklace models explicitly define the volume of each monomer and represent it as a sphere. The linear pearl necklace model describes the spheres joined by a rigid rod. The scattering function for this model (Dobrynin et al., 1996) does not include explicit terms for the scattering contribution of the rigid rod or the scattering interference between the spheres and the rigid rod. Thus, the rigid strings connecting the pearls are "invisible" in the scattering function. The parameters in this model are the radius of the spheres and the center-to-center distance between spheres. The random flight pearl necklace model describes spherical monomers connected by freely jointed rods; in other words, the spatial arrangement of the spheres is that of a freely jointed polymer chain. Because the spheres are joined to one another by a freely jointed chain, there are no excluded volume constraints on this model. In other words, two spheres may occupy the same volume. The scattering function for this model (Schweins and

Huber, 2004) includes explicit terms for the scattering from the spheres, scattering from the rods, and a cross-term for the scattering interference between the spheres and rods. The fitted-for parameters in this model are the radius of the spheres and the center-to-center distance between spheres. As an addition to these implementations, we have developed a PNM scattering function that represents the spatial arrangement of spheres as a swollen Gaussian coil and explicitly includes terms for the linker and sphere-to-linker scattering. (See Section 4.5). This model will be referred to hereafter as the excluded volume pearl necklace (EV-PNM) model. The fitted-for parameters in this model are the radius of the spheres, the persistence length of the chain, and the Flory coefficient.

In order to select the most appropriate PNM to determine the statistical conformation of 5-BdpA, we performed a weighted non-linear least squares fit of the 5-BdpA SAXS data using the scattering functions described above. The adjustable parameters for each scattering function were constrained to be non-negative and non-zero. The results are shown in Figure 25A. The best-fit parameters for each model and a measure of the goodness of fit of the scattering curves to the experimental data are given in Table 3. In the low q region (0.013 - 0.04 Å^{-1}) three models fit the data equally well: the swollen Gaussian coil model (Hammouda, 1993), the linear PNM (Dobrynin et al., 1996), and the EV-PNM. The random flight PNM (Schweins and Huber, 2004) does not fit the data. At $q > 0.04$ Å^{-1}, the best fit model to the scattering curve is the EV-PNM. This result suggests that the EV-PNM model is the most appropriate model to use when modeling the N-BdpA and SpA-N statistical conformations.

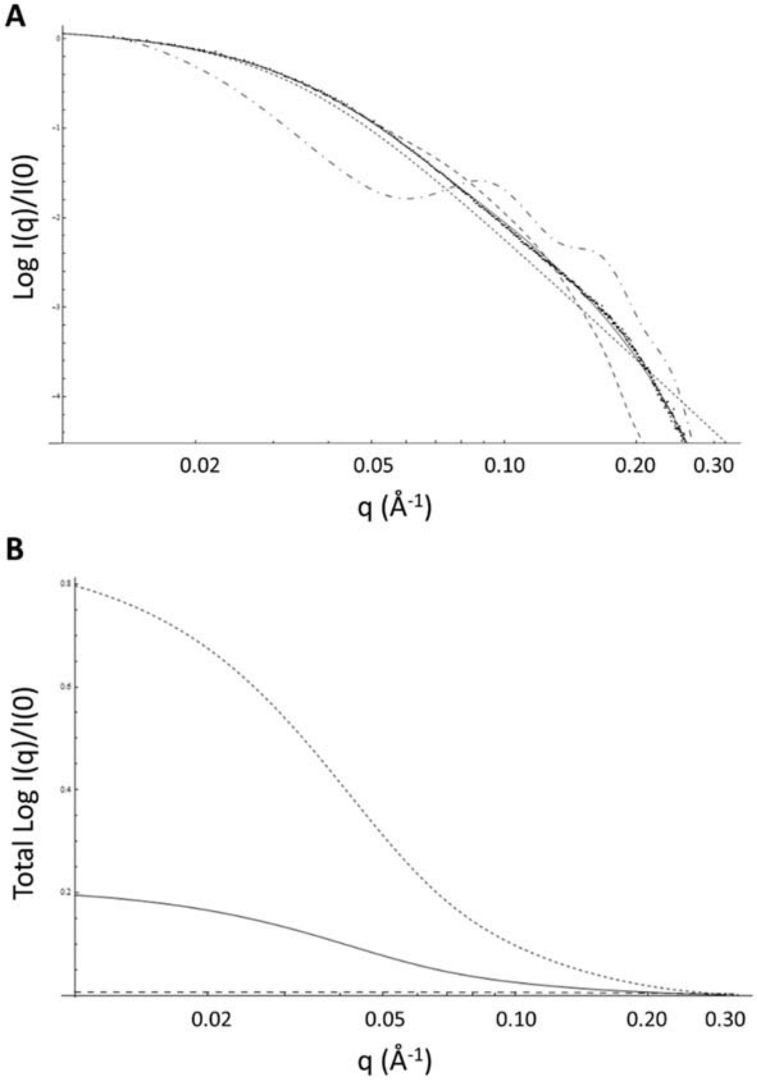

Figure 25: Polymer model comparison. A) Fit of 5-BdpA to polymer models. Points: SAXS data. Short dash: Swollen Gaussian coil model. Long dash: Linear pearl necklace model. Dash-dot: Random flight pearl necklace model. Solid line: Excluded volume pearl necklace model. For goodness-of-fit statistics see Table 3. B) Plot of I(q) vs. q showing the component scattering functions of $I_{EV\text{-}PNM}(q)$. Short dash: sphere-sphere scattering and interference. Solid line: sphere-coil interference. Long dash: coil scattering. $I_{EV\text{-}PNM}(q)$ has been scaled to $I_{EV\text{-}PNM}(0)=1$.

Table 3: Fit of the 5-BdpA SAXS data to polymer models

Polymer Model	Parameters	95% Confidence Intervals
Swollen Gaussian Coil	$\chi^2 = 2.48$ R_g: 45.00 Å ν: 0.68	 44.30 – 45.67 Å 0.67 – 0.69
Linear Pearl Necklace	$\chi^2 = 2.28$ Distance between spheres: 28.32 Å Radius: 15.00 Å	 25.10 – 31.35 Å 15.97 – 16.00 Å
Random Flight Pearl Necklace	$\chi^2 = 3.00$ Distance between spheres: 57.13 Å Radius: 12.81 Å	 54.02 – 60.23 Å 12.70 – 12.91 Å
Excluded Volume Pearl Necklace	$\chi^2 = 1.06$ Persistence length: 36.30 Å ν: 0.76 Radius: 10.5 Å	 31.3 – 37.4 Å 0.72 – 0.80 9.9 – 11.4 Å

There are two simplifications in the scattering function derived from the EV-PNM model: the shape of the domains is approximated as a sphere and the structure factor for coil to coil interference is not included. A spherical form factor describing the shape of BdpA is a reasonable and minimally parameterized model for each domain. A more complicated model for the domains would require more parameters in the scattering function, which would not be supported by the data because the fit is excellent without these parameters. The small difference in the data vs. fit at high-q could be due to the non-spherical shape of each domain. The scattering amplitude of each sphere is 8.17 times larger than the scattering amplitude of each coil, so the coil-to-coil interference is a very small component of the total scattered intensity and can be neglected (Figure 25B).

The good fit of the 5-BdpA scattering data to the EV-PNM scattering function suggests that the scattering data from the N-BdpA and SpA-N protein fragments can be globally fit to the model, resulting in a comprehensive description of the conformational space of SpA-N.

4.2.3 Global fit of the EV-PNM to the scattering data

A global fit of the scattering data from 3-BdpA, 4-BdpA, 5-BdpA, ((3-5)-BdpA) and SpA-N protein constructs to the EV-PNM scattering function using a weighted non-linear least squares fitting algorithm gives a monomer radius of 11.1 Å (95.4% confidence intervals: 10.0 - 11.8 Å), a persistence length of 35.6 Å (95.4% confidence intervals: 30.2 - 38.0 Å), and a Flory coefficient of 0.75 (95.4% confidence intervals: 0.72 - 0.80). The persistence length (37.5 ± 1.3 Å) and Flory coefficient (0.68 ± 0.04) determined using the R_g data are within the 95.4% confidence intervals determined by the global fit of all the scattering data. The 2-BdpA SAXS data was not fit to the EV-PNM, but was fit to the barbell model (see below).

The fits of the data to the EV-PNM and the goodness of fit statistics are presented in Figure 26. The model captures the dominant features of the scattering data in the (3-5)-BdpA and SpA-N constructs. The best fit of the data to the model is in the 5-BdpA data (Fig. 26d, $\chi^2 = 1.06$). The worst fit of the data to the model is 3-BdpA (Fig. 26a, $\chi^2 = 1.18$). The discrepancy may be a consequence of the low number of monomers (N=3), which may limit the ability of the EV-PNM to depict the statistical conformation of 3-BdpA. There is good agreement between the model and data at 0.013 Å$^{-1} < q < 0.1$ Å^{-1}, indicating that the model adequately describes the statistical ensemble at large length-scales (62 - 500 Å). There is a systematic deviation between the model and the data over the range 0.1 Å$^{-1} < q < 0.32$ Å^{-1}. This q region corresponds to real-space dimensions of 20 - 62 Å. In this region, deviations in local structure from the model can account for the discrepancy between the model and data, particularly deviations in the shape of the monomers. The BdpA domains are not spherical (discussed below). Our simple EV-PNM model does not take this refinement into consideration. Detailed information about the structure of the monomer units is outside the scope of these experiments because their purpose is to determine the statistical conformation of SpA-N. Simplifying the shape of the domains to a sphere is sufficient to determine the overall shape of SpA-N and (3-5)-BdpA. This minimally parameterized model describes the arrangement of the domains relative to each other at a level consistent with the information content of the data.

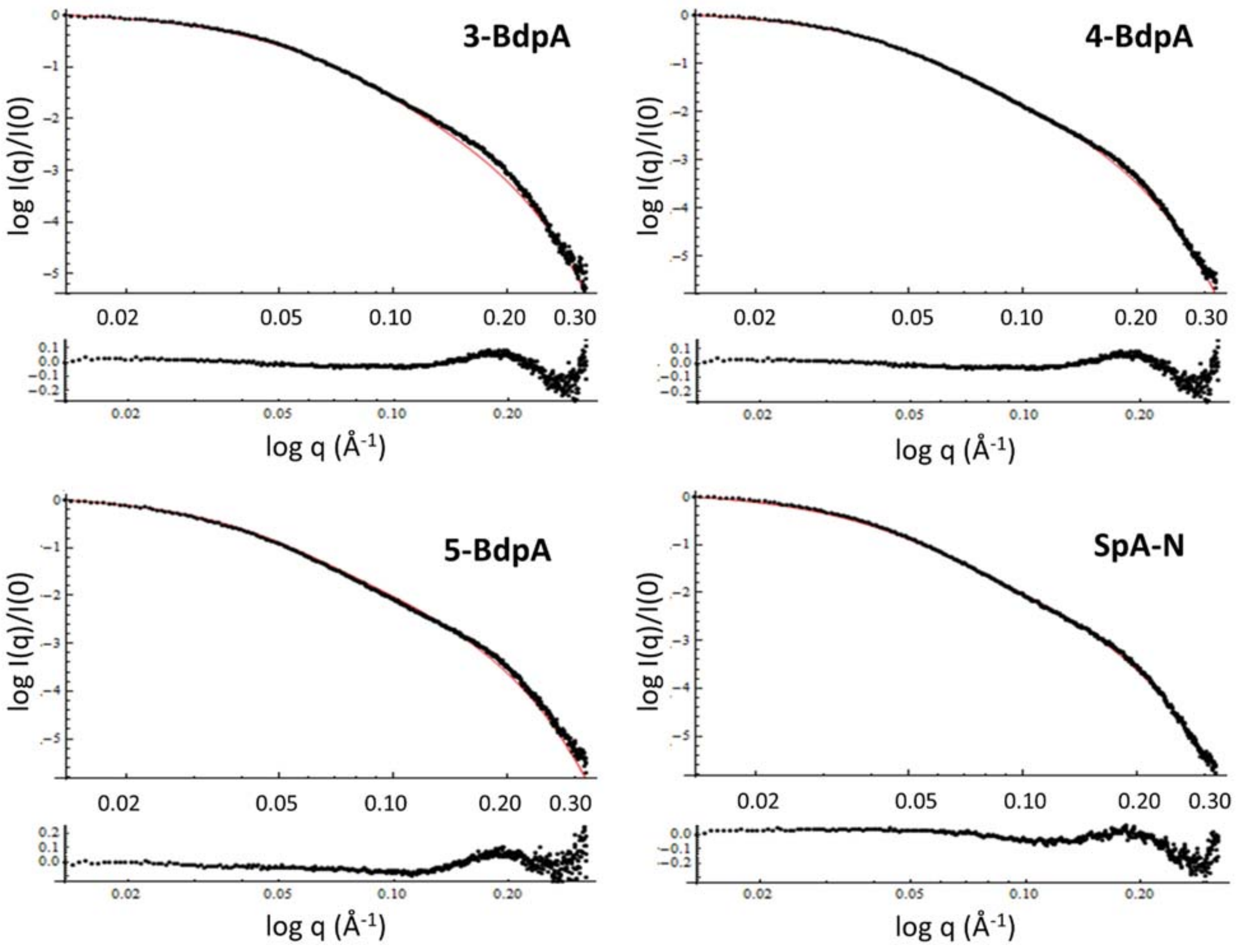

Figure 26: Global fit of the EV-PNM model to the (3-5)-BdpA and SpA-N SAXS data. Black: data. Red: model. Bottom panel of each plot: residuals (data-model). The χ^2 statistics for the global fit of the model to each dataset are as follows: 3-BdpA: 1.18, 4-BdpA: 1.12, 5-BdpA: 1.06, SpA-N: 1.07

The best-fit radius of the monomer sphere is similar to the radius of a single BdpA domain (Figure 27). Figure 27 shows the solution structure of BdpA (grey) and a sphere of 11.1 Å centered on its center of mass (black). Most of the mass of the globular portion of BdpA is contained within the sphere. There is volume in the sphere that is not occupied by BdpA but some of the side-chain volume is outside the modeled sphere, which partially compensates for the empty space. Also, as expected, the mass of the linker residues is outside the sphere. These discrepancies confirm that a sphere is not a perfect model for an individual domain. Despite deviations in the atomic details, the agreement between the EV-PNM overall domain dimensions and those from the structure of BdpA supports the accuracy of the persistence length and Flory coefficient we obtained from the global fit of the (3-5)-BdpA data.

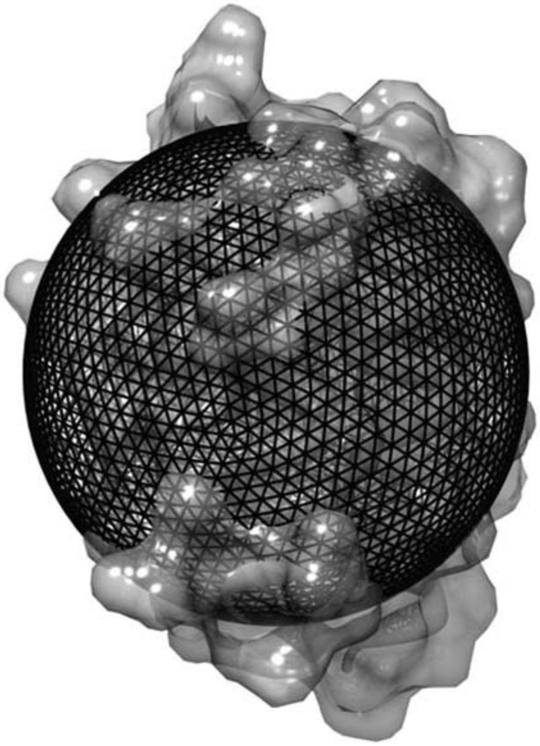

Figure 27: Space filling model of BdpA (grey), pdb: 1Q2N, superimposed with a sphere of 11.05 Å (black). The linker residues have been trimmed from the space-filling model.

4.2.4 Calculation of the end-to-end distance distribution of (3-5)-BdpA

Using the sphere radius, persistence length, and Flory coefficient determined by the fit of the SAXS data to the EV-PNM, it is possible to calculate an end-to-end distance distribution for (3-5)-BdpA. Using polymer theory, the root-mean-squared end-to-end distance ($\langle R_{ETE}^2 \rangle$) of an excluded volume Gaussian coil is (Kurata et al., 1958):

$$\langle R_{ETE}^2 \rangle = N\left(\frac{l_p}{2}\right)^2 \left(1 + \frac{4\left(\frac{3}{2\pi\left(\frac{l_p}{2}\right)^2}\right)^{\frac{3}{2}} \beta N^{\nu}}{3} \right) \qquad 4.2$$

where $\beta = \frac{4}{3}\pi R^3$ is the excluded volume and is dependent on the radius, R, of the spherical domains (determined above), and N is the number of monomers. The end-to-end distance distribution, P(R_{ETE}), of a pearl-necklace polymer is (Kamide and Dobashi, 2000);

$$\mathrm{P}(R_{ETE}) = \left(\frac{3\left(\frac{R_{ETE}}{\sqrt{\langle R_{ETE}^2 \rangle}}\right)}{2\pi} \right)^{\frac{3}{2}} e^{\left(\frac{-3R_{ETE}}{2\sqrt{\langle R_{ETE}^2 \rangle}}\right)^{\frac{1}{(1-\nu)}}} \qquad 4.3$$

Figure 28 shows the two- and three-dimensional plots of the normalized end-to-end distance distribution (EEDD) for (3-5)-BdpA. The root-mean-squared end-to-end distance for 3-BdpA is 43.6 Å, 53.4 Å for 4-BdpA, and 62.2 Å for 5-BdpA. The hemispherical representation of the EEDD is meant to represent the probability of finding the N-terminal domain at some point in space, assuming that the center of the C-terminal domain coincides with the origin of the hemisphere. This representation is analogous to the presentation of SpA-N on the surface of an *S. aureus* cell, to which the protein is attached at its C-terminus (Guss et al., 1984) . Note the low probability near the origin, which reflects the excluded volume feature of the model. Also note that the radius of maximum probability does not increase linearly with the number of domains, which it would if the molecule were rigid and linear.

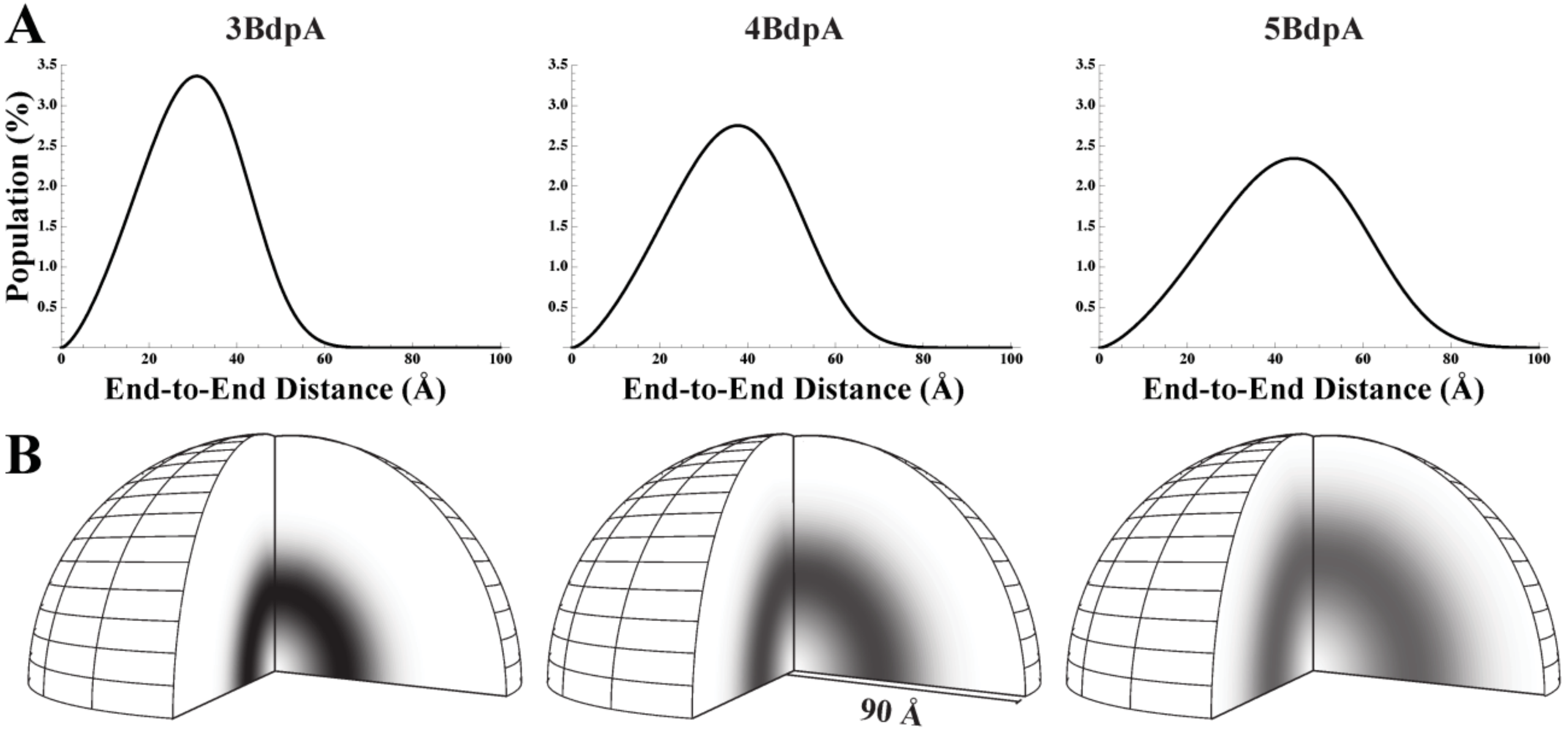

Figure 28: The end-to-end distance distribution (EEDD) of (3-5)-BdpA. A) The two dimensional EEDD for (3-5)-BdpA. B) Three-dimensional EEDD for (3-5)-BdpA, assuming that the C-terminal domain is tethered to a surface (e.g. cell wall) and is at the center of the hemisphere.

4.2.5 Fit of 2-BdpA to the barbell model

The scattering data from 2-BdpA cannot be modeled by the EV-PNM since this model assumes N ≥ 3. This data can, however, be fit to a modified barbell model (BM). This model describes two spheres separated by a line with no scattering mass. Our variation of the scattering model for this model explicitly includes a term for the scattering of the linker (see Section 4.5). We fit the 2-BdpA scattering data to this model and obtained an average distance between domains of 5 Å (± 0.24 Å) and a radius of 15.32 Å (± 0.10 Å). The radius of the spheres is larger than that determined by the global fit of (3-5)-BdpA and SpA-N to the EV-PNM, but it is still reasonable given the structure of BdpA.

4.3 Discussion

4.3.1 Model of the SpA-N statistical conformation

The persistence length of a polymer can be used to describe the length scale of polymer flexibility. The statistical conformation of a polymer at length scales smaller than the persistence length is rigid, while at length scales larger than the persistence length it is flexible (Fujita, 1990). The persistence length obtained from the global fit of (3-5)-BdpA and SpA-N (35.6 Å) is similar to the length of a single BdpA domain (29.6 Å). This result indicates that SpA-N is a highly flexible biopolymer. To gain a more intuitive understanding of *how* flexible the biopolymer is, we can compare the persistence length of (3-5)-BdpA and SpA-N to the persistence lengths of other well- known polymers and biopolymers (Table 4). The persistence length of an unfolded polypeptide, (6.6 Å) (Lairez et al., 2003), is an order of magnitude smaller than that of SpA-N. This result is expected since SpA-N is not an unfolded protein, but rather is composed of 5 globular domains. Surprisingly, the persistence length of SpA-N is quite similar to that of ssDNA (22.2 Å) (Chi et al., 2013), even though the monomer dimensions are very different.

Table 4: The persistence length of various polymers

	SpA-N	Unfolded Polypeptide[1]	Polystyrene[2]	ssDNA[3]	dsDNA[4]
Persistence Length	35.6 Å	6.6 Å	10 Å	22.2 Å	500 Å

[1](Lairez et al., 2003), [2](Wignall et al., 1974), [3](Chi et al., 2013), [4](Hagerman, 1988)

We must ask: what are the physical and chemical properties of SpA-N that confer this flexibility? The excluded volume constraint imposed by the dimensions of the monomers lends stiffness to the chain. The peptide bond geometry, sterics, and solvent-peptide interactions constrain the statistical conformation of the linker lending stiffness to the chain. However, if there were inter-domain attraction or if the linker was completely rigid, the chain would have a much larger persistence length and much larger chain dimensions. We therefore hypothesize that the only constraints on the statistical conformation of SpA-N are excluded volume interactions between domains and the constraints imposed by the chemical properties of the linker. This hypothesis is consistent with our previous research on N-BdpA and SpA-N. Careful denaturation experiments comparing the stability of each isolated domain and the same domain within the SpA-N molecule showed that the folding of SpA-N domains is thermodynamically uncoupled. Similar studies showed that N-BdpA has the same denaturation curve for n = 1 - 5, again demonstrating the lack of thermodynamic interaction between domains. NMR relaxation studies of backbone ^{15}N-^{1}H pairs showed that the order parameters of all residues in 5-BdpA are high except for those in the termini and a 6-residue linker between each domain. These results demonstrated that the linker residues are almost as flexible as the corresponding residues in the termini. Taken together, both previous results and the present SAXS data strongly support the conclusion that SpA-N and N-BdpA are highly flexible chains of inflexible domains that lack any significant favorable inter-domain interactions.

Describing the statistical conformation of SpA-N and N-BdpA with a polymer physics model provides a continuous description of conformational space - we do not discretize conformational space into an ensemble of unique conformers. Based on the SAXS data presented above, the global conformational space of SpA-N includes the fully extended conformations and quite compact conformations and all conformations between these two extremes, as long as they avoid steric overlap. The SpA-N SAXS data is inconsistent with a single compact conformation, or a thermal blob (multi-conformation compact conformers), or an elongated conformation. Instead, our EV-

PNM fit of the SAXS data implies that the statistical conformation includes all these conformations and intermediate conformers as well.

It is important to note that this continuous description of the statistical conformation is consistent with our previous knowledge of the structure of SpA-N and is consistent with the information content of the SAXS data. Describing the statistical conformation of SpA-N to any higher "resolution" would over-interpret the SAXS data and over-parameterize the model of allowed SpA-N conformations. Our SAXS data cannot provide us with any information about the atomistic detail of each domain or linker, cannot help us determine a "minimal ensemble" of SpA-N conformers (see below) and cannot provide information about any anisotropic domain-domain motion in the statistical conformation. It may be that such anisotropies exist because of the non-spherical shape of the domains, but given the excellent fits to the simplistic EV-PNM model, we conclude that the SAXS data does not contain any information regarding such anisotropies. Other biophysical techniques are needed to further limit allowed conformational space. Most importantly, the SAXS data cannot provide us with any information about the distribution of conformational space, since SAXS data is a population-weighted average of all allowed conformational space. The continuous statistical description of conformational space provided by the EV-PNM most accurately describes the allowed conformational space of SpA-N and is consistent with the information content of the SAXS data. Therefore, we do not present a "structure" of the statistical conformation of SpA-N beyond the end-to-end distance distribution, persistence length, Flory coefficient, and radius of the identical spheres that represent the domains.

4.3.2 Ensemble modeling of SpA

In order to compare the description of conformational space resulting from the fit of the EV-PNM to our SAXS data to the description of conformational space resulting from ensemble based methods, I used an ensemble modeling method to analyze the 2-BdpA and 5-BdpA SAXS data. Following the EOM protocol (Bernado et al., 2007), I used a modified RanCH algorithm that produces a structurally converged parent ensemble of atomic-resolution structures to generate an unconstrained parent ensemble of self-avoiding conformations. I then ran the GAJOE algorithm multiple times to select from this parent ensemble 14-50 conformations (minimal ensemble) whose calculated aggregate SAXS curve matched the observed data. I observed that, for 5-BdpA, the distribution of global structural parameters, R_g and D_{max}, of the minimal ensemble was statistically identical to the distribution of the parent ensemble (Figure 29B). In the case of 2-BdpA, the best fit minimal ensemble was bi-modally or tri-modally partitioned into more compact and more extended conformations (Figure 29A). However, there was no agreement of the minimal ensemble between successive runs of GAJOE. In all instances, the fit of the minimal ensemble to the SAXS data was nearly identical. Most significantly, in the case of both 2-BdpA and 5-Bdpa, the calculated aggregate SAXS curve of the unconstrained parent ensemble matched the data nearly as well as any of the minimal ensemble curves (Figure 30). This result indicates that, in terms of a discrete atomistic model, no information about the statistical conformation can be obtained from the SAXS data other than it is consistent with a self-avoiding chain of domains linked via six-residue flexible linkers.

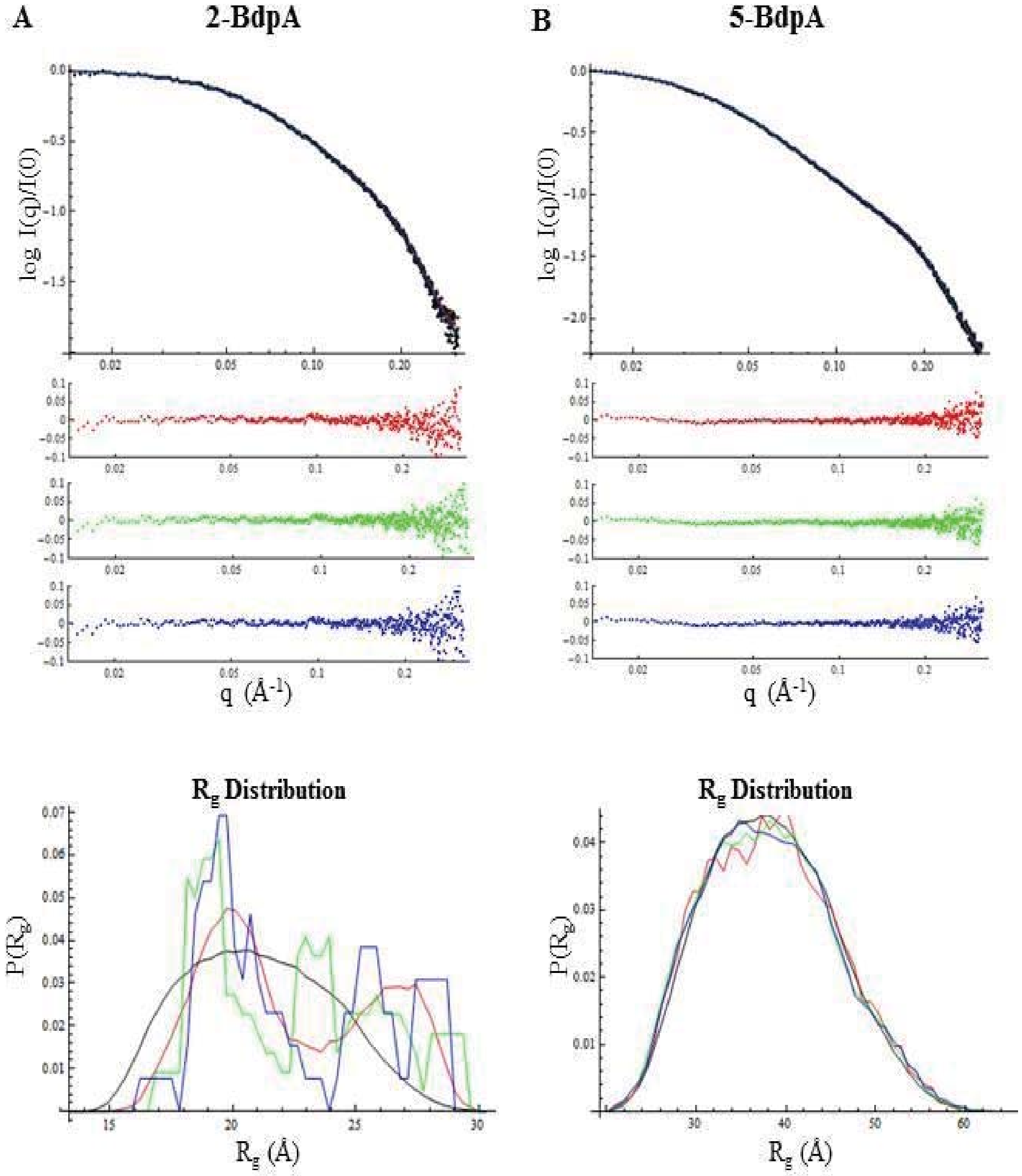

Figure 29: Ensemble analysis of 2-BdpA and 5-BdpA. A) The fit of the 2-BdpA scattering curve from the minimal ensembles to the 2-BdpA data (black), with residuals (data-model) below the data (red, green, blue), the χ^2 statistic (calculated by EOM) is 1.303 for all three ensembles. Bottom panel: R_g distribution of the parent ensemble (black) and each of the three calculated minimal ensembles (red, green, blue). Note that although the fit of the calculated ensemble scattering curve is

indistinguishable between each of the three minimal ensembles selected, the R_g distribution for each ensemble varies widely. B) The fit of the 5-BdpA calculated scattering curve from the minimal ensemble to the 5-BdpA(black) data. Residuals are plotted below the scattering curves (red, green, blue). The χ^2 statistic (calculated by EOM) for each ensemble is: 1.146 (red and green) and 1.147 (blue). Bottom panel: R_g distribution of the parent ensemble (black) and each of the three calculated minimal ensembles (red, green, blue). Note that the R_g distributions of the minimal ensemble are statistically similar to the R_g distributions of the parent ensemble.

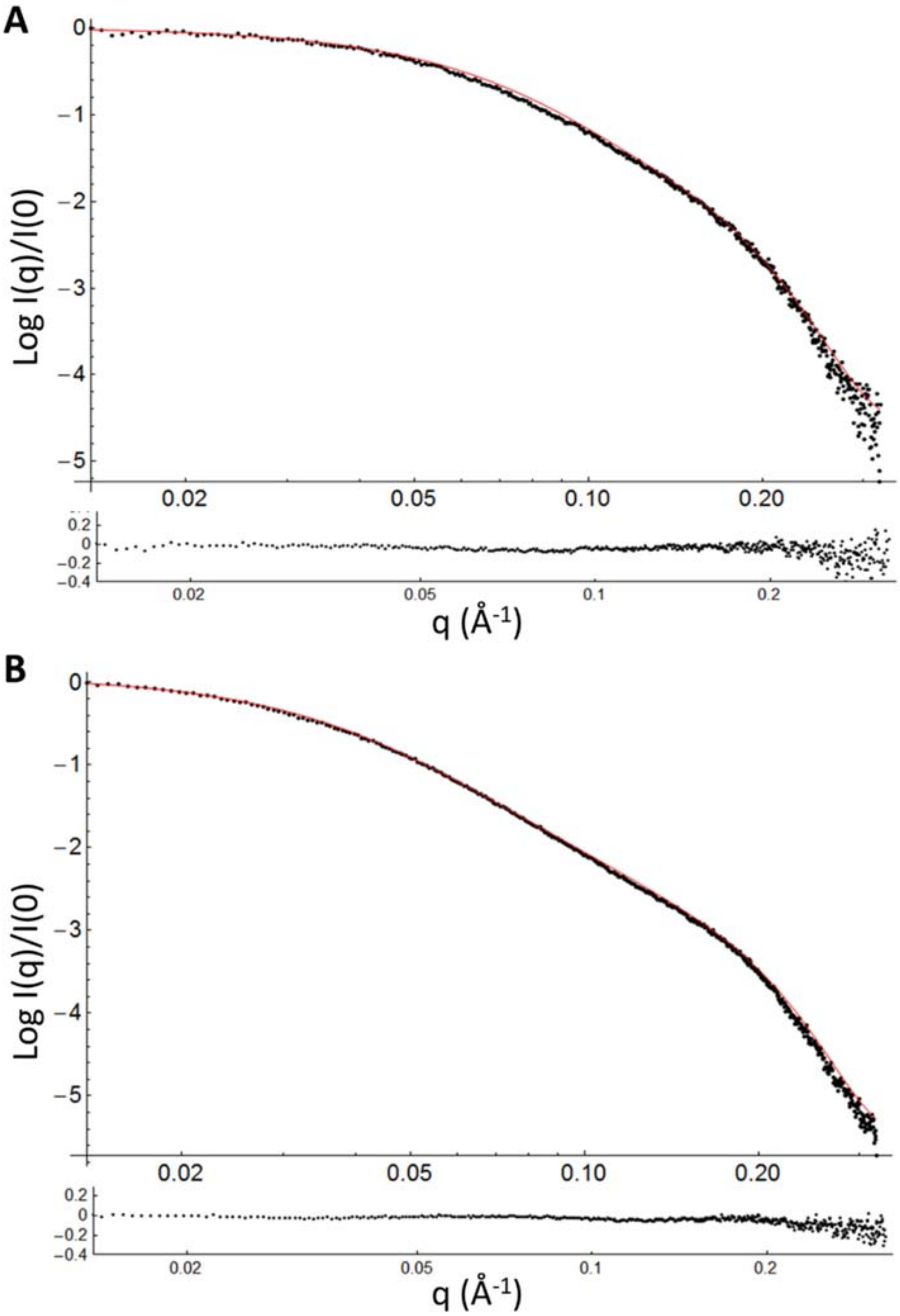

Figure 30: The aggregate scattering curves calculated from the unconstrained parent ensembles fit the experimental data very well. CRYSOL (Svergun et al., 1995) was used to calculate the theoretical scattering curve for each structure in the

unconstrained parent ensemble. The aggregate scattering curve was calculated by summing the scattering curves for each structure in the parent ensemble. The aggregate scattering curve was normalized by I(q)/I(0). (A) Fit of the 2-BdpA aggregate scattering curve from the unconstrained parent ensemble (red) to the 2-BdpA experimental data (black), $\chi^2 = 1.04$. Residuals (data-model) are plotted below the scattering curve. (B) Fit of the 5-BdpA aggregate scattering curve from the unconstrained parent ensemble (red) to the 5-BdpA experimental data (black), $\chi^2 = 1.04$. Residuals (data-model) are plotted below the scattering curve.

4.3.3 The structural flexibility of SpA may contribute to its functional plasticity and allow for maximum binding of ligands in the extracellular environment

An evolving view of the protein structure-function relationships is that conformationally dynamic proteins can exhibit functional promiscuity (Tokuriki and Tawfik, 2009). Some highly flexible proteins can recognize multiple ligands at a single binding surface. This structural flexibility allows for the accommodation of mutations as two proteins co-evolve and allows a single molecule to interact with multiple structurally unique binding partners.

SpA exhibits both structural flexibility and functional plasticity. Its sequence has evolved to perform a wide array of functions that confer virulence to *S. aureus* including binding to both Fc and Fab fragments of antibodies (Deisenhofer, 1981; Moks et al., 1986); von Willebrand factor (Hartleib et al., 2000); and TNFα receptor (Gomez et al., 2004). This panoply of binding partners no doubt requires corresponding structural plasticity on the part of SpA. One manifestation of this structural plasticity might be the flexibility we observe between domains. Because the SpA statistical conformation includes a large ensemble with a variety of inter-domain orientations, it can accommodate multiple ligands binding in many different contexts. In principle this would give it the potential to rapidly evolve in response to changing environmental conditions, conferring resistance and adaptability to *S. aureus.*

The high flexibility of SpA assures that the surface available for interaction with cognate binding partners around each SpA attachment in the cell wall is maximized. The high abundance of SpA in the *S. aureus* cell wall (Sjoholm et al., 1972) suggests that, in aggregate, this interaction surface could represent a large fraction of the bacterial surface. This high surface availability may be a key determinant of SpA's function as a virulence factor: both the abundance of SpA and it's flexibility would maximize SpA:ligand interaction.

Because SpA flexibility is the consequence of a short, conserved six amino acid segment between domains, it would be feasible to fully explore the sequence-dependence of flexibility. If such studies were to yield a set of sequences with a wide range of flexibilities, it would be possible to directly test the biological significance of SpA flexibility. It is possible that a change in the flexibility of the linker could result in a change in affinity for some of SpA's ligands.

4.4 Concluding Remarks

In this chapter, SAXS was used to determine and describe the statistical conformation of *S. aureus* protein A. It was demonstrated that SpA is a highly flexible protein, and the only constraints on the statistical conformation, detectable by SAXS, are excluded volume constraints. In the following chapter I will show how a SAXS analysis of another highly flexible protein can help resolve a conflict between two different models that describe its structure and function.

4.5 Deriving the BM and EV-PNM scattering functions:

4.5.1 The Barbell Model (BM): Introduction

2-BdpA can be modeled as a polymer composed of two phases: homogeneous domains and flexible linkers. The two domains are described as spheres with a form factor, F_{sphere}. The structure factor (S_{lin}) describes the interference due to the two spheres separated by a fixed distance. S_{lin} can be fit to an experimental SAXS profile to determine the average distance between the two domains. The flexible regions of the protein (N-terminus and inter-domain linker) are modeled as excluded volume Gaussian coils. This form factor is denoted F_{ev}. The analytical derivations of F_{sphere}, F_{ev}, and S_{lin} are described below. Substituting these form and structure factors into Eq 2.10, gives the following result for the sums over all components:

$$I_{BM}(q) = 2\,F_{sphere}^2(q)\,S_{lin}(q) + 4\,F_{sphere}(q)\,F_{ev}(q)\,S_{lin}(q) + 2\,F_{ev}^2(q) \qquad 4.4$$

In $\mathrm{I_{BM}(q)}$ the length of the linkers and the distance between two domains is uncorrelated. Note: there is no function for the scattering interference between the linkers (see section 4.2.2). The first term of Eq. 4.4 accounts for scattering of the spheres and the interference between them; the third term accounts for coil scattering, and the middle term accounts for coil-sphere interference. Eq. 4.4 ignores interference between coils because this term represents a very small fraction of the total scattering of 2-BdpA.

4.5.2 Scattering function: Barbell model

$\mathrm{I_{BM}(q)}$ can be decomposed into three components: scattering and interference from spheres ($\mathrm{F^2_{sphere}(q)\ S_{ev}(q)}$), scattering from coils ($F^2_{ev}(q)$), and a cross term ($2\ F_{sphere}(q)\ F_{ev}(q)\ S_{ev}(q)$), that accounts for correlations between interconnecting spheres and coils. The so-called "coils" of the model are equivalent to the inter-domain linker and the 6 residues at the termini.

4.5.2.1 Scattering from domain spheres: $\boldsymbol{F^2_{sphere}(q)\ S_{BM}(q)}$.

The normalized scattering amplitude of the domains is assumed to be the form factor of a sphere (Glatter and Kratky, 1982):

$$\mathrm{F_{spheres}(q)} = 3\frac{\sin(qR)-(qR)\cos(qR)}{(qR)^3}\ \mathrm{W_s} \qquad 4.5$$

R is the radius of the spheres, and W_s is the ratio of the scattering mass of the domains to the scattering mass of the coil. For 2-BdpA, this ratio has been determined by

NMR dynamics data. The coils consist of 12 (6 residues each) residues, and the spheres consist of 52 residues each. This gives a scattering mass ratio of 11647.2:1425.6 = 8.17.

The structure factor describing spatial arrangement of the spheres in the biopolymer is (Schweins and Huber, 2004):

$$\mathrm{S_{BM}(q)} = \frac{2}{1-\frac{\sin(qb)}{qb}} - 1 - \frac{1-\left(\frac{\sin(qb)}{qb}\right)^2}{\left(1-\frac{\sin(qb)}{qb}\right)^2} * \frac{\sin(qb)}{qb} \qquad 4.6$$

where b is the distance between the domain spheres, the only fitted parameter in this term.

4.5.2.2 Scattering from the coils: $F_{ev}^{2}(q)$

As Hammouda notes (Hammouda, 1993), the square of the form factor for an excluded volume Gaussian coil is not simply $F_{ev}^{2}(q)$, but is instead:

$$\mathrm{F_{ev}^2}(q) = S_{ev}(q) = \frac{1}{vX^{\frac{1}{2}}} * \Gamma(\frac{1}{2v}, X) - \frac{1}{X^{\frac{1}{2v}}} * \Gamma(\frac{1}{v}, X) \qquad 4.7$$

where $\Gamma(a, X)$and X are:

$$\Gamma(a, X) = \int_0^X dt\ exp(-t) t^{a-1}, \text{ and} \qquad 4.8$$

$$\mathrm{X} = \frac{\mathrm{q^2 (\frac{lp}{2}) N^{2v}}}{6} \qquad 4.9$$

In the case of the 2-BdpA linker, v = 0.588 (the Flory coefficient for unfolded peptides) (Baldwin, 2002), l_P = 6.6 (the persistence length of unfolded proteins) (Lairez et al., 2003), and N = 6 (the number of residues in the linker). There are no adjustable parameters in this term.

4.5.2.3 Cross-terms: 2 $\boldsymbol{F_{sphere}(q)\,F_{ev}(q)\,S_{BM}(q)}$

The components $F_{sphere}(q)$and $S_{BM}(q)$ have been defined above. In this case, $S_{ev}(q)$refers to the spatial arrangement of the spheres. $F_{ev}(q)$is the form factor for the coils and is (Hammouda, 1993):

$$F_{ev}(q) = \frac{1}{vX^{\frac{1}{2v}}} * \Gamma\left(\frac{1}{2v}, X\right) - \frac{1}{vX^{\frac{1}{v}}} * \Gamma\left(\frac{1}{v}, X\right) \qquad 4.10$$

The fitted parameters in this term are R (the radius of the domain spheres) and b (the distance between spheres).

4.5.3 The Excluded Volume Pearl-Necklace Model (EV-PNM): Introduction

N-BdpA can be similarly modeled using an excluded volume pearl-necklace model (EV-PNM) for n > 2. The two phases of this model are made up of n homogeneous domains and n homogenous flexible regions of the protein (N-terminus and inter-domain linkers). The domains are modelled as spheres with a form factor, F_{sphere}. The spatial arrangement of domain spheres is represented by the structure factor for an excluded volume Gaussian coil, S_{ev}. S_{ev} includes parameters that model the flexibility of the chain of spheres: the persistence length (l_p) and the Flory coefficient (ν). The model for the flexible regions is identical to the description in the BM. Substituting these form and structure factors into equation 2.10 gives the following results for the sums over all components:

$$\mathrm{I_{EVPNM}(q) = \ NF^2_{sphere}(q)\,S_{ev}(q)\ +\ 2\,NF_{sphere}(q)\,F_{ev}(q)\,S_{ev}(q)\ +\ NF^2_{ev}(q)} \qquad 4.11$$

N is the number of BdpA domains in the protein. For SpA-N, N=5. In equation 4.11 the length of the linkers and the distance between two domains is uncorrelated. Equation 4.11 also ignores interference between coils because this term represents a very small fraction of the total scattering of N-BdpA.

4.5.4 Scattering function - excluded volume pearl necklace model:

$\mathrm{I_{EVPNM}(q)}$ can be decompose into three components: scattering and interference from spheres ($\mathrm{F^2_{sphere}(q)\ S_{ev}(q)}$), scattering from coils ($F^2_{ev}(q)$), and a cross term ($2\ F_{sphere}(q)\ F_{ev}(q)\ S_{ev}(q)$), accounting for correlations between interconnecting spheres and coils.

4.5.4.1 Scattering from domain spheres: $\boldsymbol{F^2_{sphere}(q)\ S_{ev}(q)}$

The normalized scattering amplitude of the domains is assumed to be sphere, identical to that in the Barbell model:

$$\mathrm{F_{spheres}}(q) = 3\,\frac{\sin(qR)-(qR)\cos(qR)}{(qR)^3}\ \mathrm{W_s} \qquad 4.12$$

R is the radius of the spheres, and W_s is the ratio of the scattering mass of the domains to the scattering mass of the coil. For SpA-N and N-BdpA, this ratio is 8.17/1.

The structure factor describing the interference due to the spatial arrangement of spheres within the biopolymer is (Hammouda, 1993):

$$S_{ev}(q) = \frac{1}{vX^{\frac{1}{2}}} * \Gamma(\frac{1}{2v}, X) - \frac{1}{X^{\frac{1}{2v}}} * \Gamma(\frac{1}{v}, X) \qquad 4.13$$

where $\Gamma(a, X)$ is the incomplete gamma function:

$$\Gamma(a, X) = \int_0^X dt \exp(-t)\, t^{a-1}, \tag{4.14}$$

and X is:

$$X = \frac{q^2 (\frac{lp}{2}) N^{2v}}{6}. \tag{4.15}$$

v is the Flory coefficient, l_P is the persistence length, and N is the number of domains. The adjustable parameters in $F_{sphere}^2(q)\, S_{ev}(q)$are R, v, and l_P.

4.5.4.2 Scattering from the coils: $\boldsymbol{F_{ev}^2(q)}$

This term is identical to that of the Barbell model, Eq.4.7.

4.5.4.3 Cross-terms: 2 $\boldsymbol{F_{sphere}(q)\, F_{ev}(q)\, S_{ev}(q)}$

The components $F_{sphere}(q)$and $S_{ev}(q)$ has been defined above. In this case, $S_{ev}(q)$refers to the spatial arrangement of the spheres. $F_{ev}(q)$is the form factor for the coils and is (Hammouda, 1993):

$$F_{ev}(q) = \frac{1}{vX^{\frac{1}{2v}}} * \Gamma\left(\frac{1}{2v}, X\right) - \frac{1}{vX^{\frac{1}{v}}} * \Gamma\left(\frac{1}{v}, X\right) \tag{4.16}$$

where $\Gamma(a, X)$and X are defined above. The fitted for parameters in this function are R (the radius of the domain spheres), v (the Flory coefficient for the sphere spatial relationship), and l_P (the persistence length of the entire chain).

4.6 Materials and Methods

4.6.1 Protein expression and purification:

Plasmid constructs were transformed into *E. coli* BL21(DE3) cells using standard procedures. 1L LB media containing 100 mg/L ampicillin was then inoculated with a single colony of the transformed cells. The cells were grown at 37° C until an OD_{600} of 0.8-1.0 was reached. The culture was induced with IPTG to a final concentration of 1mM then harvested 4-6 hours post-induction and centrifuged. The cell pellet was resuspended in 50mM Tris pH 8.8, 1mM EDTA and protease inhibitors (AEBSF, pepstatin, bestatin and E-64). The cells were lysed and insoluble material was cleared by centrifugation. The pH of the cleared lysate was adjusted to pH 9.0 and 10 μL micrococcal nuclease was added to digest large DNA fragments. The resulting solution was brought to 4M guanidinium HCl and 20mM TCEP by the addition of solid guanidinium HCl (Bio-Basic) and 1M TCEP. The solution was dialyzed into a 5% acetic acid solution, insoluble materials were cleared by centrifugation. The soluble material was dialyzed into deionized water. The protein solution was loaded onto an SP Sepharose (GE Healthcare) column in 50mM sodium acetate, pH 3.6. The protein was eluted from the column by a 600ml 100 – 500mM NaCl gradient in 8ml fractions. The fractions were checked for purity by SDS-PAGE. The most pure fractions were pooled and dialyzed against deionized water. This protein solution was loaded onto a DEAE Sephacil (GE Healthcare) column in 50mM sodium acetate, pH 3.6. The protein was eluted from the column by an 800ml 0-250 mM NaCl gradient in 8ml fractions. The fractions were checked for purity by SDS-PAGE. The most pure fraction were pooled and dialyzed into deionized water. The protein was then lyophilized and stored in a desiccator.

4.6.2 SAXS sample preparation:

Lyophilized protein was resuspended in deionized water to make stock solutions. Sodium acetate, pH 5.5, sodium chloride, and glycerol was added to each stock solution to a final concentration of 50mM sodium acetate, pH 5.5, 100 mM sodium chloride, and 1% glycerol. The protein samples were then dialyzed against 50mM sodium acetate, pH 5.5, 100mM sodium chloride and 1% glycerol for 6 hours at room temperature using a 3500 Da MWCO micro-dialysis unit (Pierce).

For ALS data collection, samples were centrifuged at 16,000 x g for 20 minutes and then the concentration of each sample was calculated by A_{280}. Samples were diluted to a concentration of 5 mg/ml, 2.5 mg/ml or 1.25 mg/ml using dialysate. The samples were stored at 4° C for no more than 24 hours.

For APS data collection, each sample was stored at 4° C for no more than 36 hours prior to data collection. Just prior to data collection each sample was centrifuged at 16,000 x g for 20 minutes and the concentration was calculated by A_{280}. Samples were diluted to a final concentration of 2 mg/ml, 1 mg/ml, or 0.5 mg/ml using dialysate.

4.6.3 SAXS data acquisition and analysis

Data were collected at beamline 12.3.1 (SIBYLS) at the Advanced Light Source, Lawrence Berkeley National Labs, and at beamline 18ID (Bio-CAT) at the Advanced Light Source, Argonne National Labs.

At the SIBYLS beamline, 25 μL of protein samples were loaded into a sample cell and then exposed for 0.5, 1, or 4 seconds at an energy of 12 kEV, with a sample-to-detector distance of 1.5 M, corresponding to a q-range of 0.01 - 0.32 Å^{-1}. Data were collected from an identical buffer sample, using dialysate from the equilibrium dialysis, for each protein sample using identical data collection conditions. All data were collected at 10° C. Beamline specific software was used to reduce the data and subtract the buffer signal to generate final scattering data for each protein sample.

At the Bio-CAT beamline, 120 μL protein samples were loaded into a sample capillary and the sample was oscillated in the beam to minimize radiation damage, such that no single protein molecule was exposed for more than 100 ms. Data were collected at an energy of 12 kEV, and a sample-to-detector distance of 2750 mm, corresponding to a q-range of 0.008 - 0.29 Å . The flux of the beam was attenuated by using 18 foil attenuators. Data from identical buffer samples from the dialysate were collected for each protein sample. All data were collected at 10° C. Data were reduced using the Nika package for Igor Pro (usaxs.xray.aps.anl.gov/staff/ilavsky/nika.html). 15 individual data-sets for each protein and buffer sample were averaged in PRIMUS (Konarev et al., 2003), and the buffer signal was subtracted from the data signal using PRIMUS to generate final scattering data for each protein sample.

Guinier analysis for each construct was performed using PRIMUS to determine the radius of gyration and I(0). Polymer models were fit to the scattering data using a nonlinear least squares fitting algorithm implemented in Mathematica 9. 95% confidence intervals and standard errors were calculated in Mathematica 9. 95.4% confidence intervals were calculated according to the method of Bevington and Robinson (Bevington and Robinson, 2003).

5. The effects of ionic strength on the statistical conformation of Fibronectin type III domains 1 and 2

5.1 Introduction

Fibronectin (FN) is an extracellular matrix (ECM) glycoprotein that plays a major role in regulating the ECM assembly process and dynamics. Fibronectin is a ubiquitous protein present in all types of tissues in all stages of development. It forms a fiber-like matrix and the development and maintenance of this matrix is essential for life. Incorrect or absent ECM assembly affects fetal development, wound healing and tumorigenesis (Singh et al., 2010). Despite the importance of fibronectin in ECM assembly, the mechanisms of fibril and matrix assembly remain poorly understood.

FN is a long multi-domain protein composed of three different types of domains: FN$^{\mathrm{I}}$, FN$^{\mathrm{II}}$, and FN$^{\mathrm{III}}$ –type domains (Kornblihtt et al., 1985; Petersen et al., 1983). A single molecule of FN contains 12 FN$^{\mathrm{I}}$ domains, 2 FN$^{\mathrm{II}}$ domains, and 15-17 FN$^{\mathrm{III}}$ domains. FN type I and II domains contain intra-domain disulfide bonds that help to stabilize the folded domains. Type III domains are 7-strand β-barrel structures that lack a stabilizing intra-domain disulfide bond (Leahy et al., 1996; Potts and Campbell, 1994).

The fibronectin matrix is composed of disulfide-bonded FN dimers that are then covalently cross-linked, by an unknown mechanism, to form the FN matrix (Vakonakis et al., 2007). Several domains are proposed to play an important role in matrix assembly. FN$^{\mathrm{I}}$ domains 1-9 (FN$^{\mathrm{I}}$(1-9)) are crucial for matrix assembly (Schwarzbauer, 1991). It has been proposed that $\mathrm{FN}^{\mathrm{III}}$ domains 1 and 2 ($\mathrm{FN}^{\mathrm{III}}$(1-2)) interact with FN^{I}(1-9), or more specifically, FN^{I}(1-5), to promote matrix assembly (Sechler et al., 2001). Unlike other $\mathrm{FN}^{\mathrm{III}}$ domains, there is a large (35-residue) linker between $\mathrm{FN}^{\mathrm{III}}$ 1 and $\mathrm{FN}^{\mathrm{III}}$ 2 (Vakonakis et al., 2007). Several structural studies have proposed that $\mathrm{FN}^{\mathrm{III}}$(1-2) form a closed compact structure and that when tension is applied, they dissociate and expose a cryptic binding site for FN^{I}(1-9) that is inaccessible when they are in a compact conformation (Karuri et al., 2009; Vakonakis et al., 2007).

A recent NMR study modeled the solution structure of $\mathrm{FN}^{\mathrm{III}}$(1-2) in a compact conformation (Figure 31A). In that structure, the two domains buried considerable surface area and there was a potential salt bridge formed between a lysine in domain 1 and an aspartic acid in domain 2 (Figure 31B). When these residues were mutated to alanine (creating the KADA mutant), the authors demonstrated that binding to FN^{I}(1-5) was greatly enhanced, suggesting that there was a cryptic binding site for FN^{I}(1-9) in the $\mathrm{FN}^{\mathrm{III}}$(1-2) interaction interface that became available for binding under tension (Vakonakis et al., 2007) .

A FRET study also investigated the structure of $\mathrm{FN}^{\mathrm{III}}$(1-2) (Karuri et al., 2009). The authors showed that when $\mathrm{FN}^{\mathrm{III}}$(1-2) was inserted between two fluorescent proteins, the FRET signal was significant, indicating that $\mathrm{FN}^{\mathrm{III}}$(1-2) primarily populates a closed compact conformation. However, when the same experiment was performed with the KADA mutant, which was thought to populate the open conformation, the FRET signal was greatly enhanced. This result suggested that somehow the protein became more compact when the salt bridge stabilizing the interaction between domains 1 and 2 was eliminated instead of less compact.

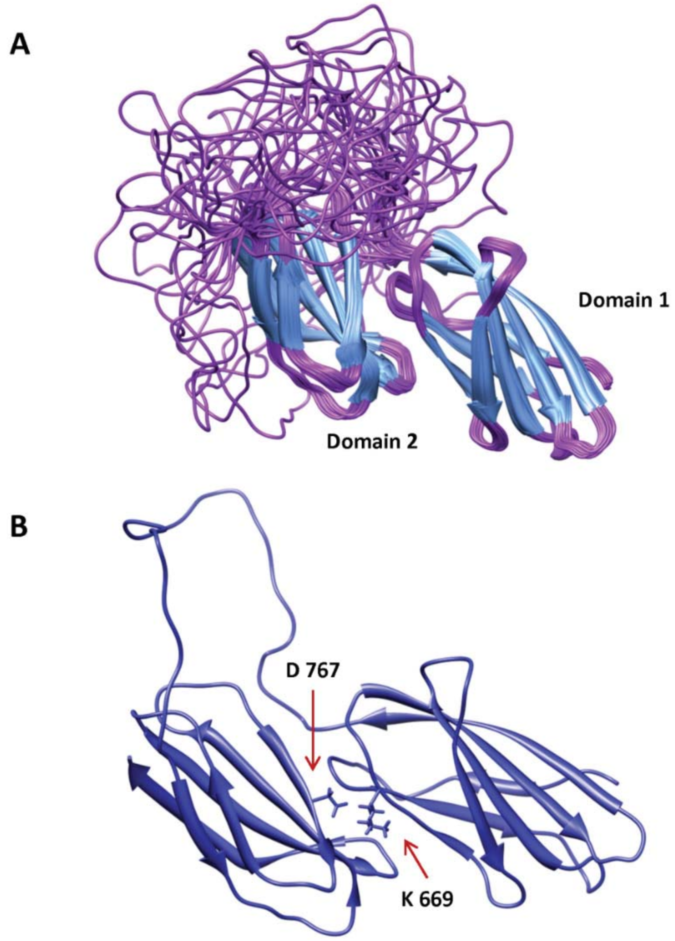

Figure 31: The NMR solution structure of FN^{III}(1-2). A) The NMR structural ensemble of FN^{III}(1-2) (pdb: 2HA1). The β-strands are colored blue, and the coils are purple. B) A single member of the NMR model (2HA1.1) with the potential salt bridge between K669 and D767 shown.

Conflictingly, extensive biophysical studies from the Erickson lab at Duke University have indicated that there is no difference in the global structural parameters of FN^{III}(1-2) and KADA (Ohashi and Erickson, 2011). Essentially, they adopt the same conformation in solution. The Erickson lab has also shown that this conformation may not be compact, but instead may be expanded. The suggested mechanism for FN^{I}(1-9) binding and fibrillogenesis resulting from these studies is a model where unfolding of FN^{III}2 is necessary for FN^{I}(1-9) binding, and the potential domain:domain contacts in FN^{III}(1-2) do not have a cryptic binding site that is accessible when tension is applied. Thus, there is currently a conflict in the literature about the statistical conformation of these two domains. A useful model of FN fibrillation and ECM formation requires additional data that resolves these conflicting descriptions of the FN^{III}(1-2) solution structure and FN^{I}(1-9) binding.

In this study, I perform a SAXS analysis of FN^{III}(1-2) in different ionic strength solution conditions. The original NMR study and Erickson lab biophysical studies were performed in the presence of 0.15M NaCl (Lemmon et al., 2011; Vakonakis et al., 2007). If the salt bridge is present and stabilizing the closed conformation in these solution conditions, then we would expect that upon an increase of ionic strength, the statistical conformation of FN^{III}(1-2) would expand. Similarly, at NaCl concentrations less than 0.15M, we would expect the statistical conformation to be nearly identical to the statistical conformation in 0.15M NaCl.

Two different FN$^{\text{III}}$(1-2) constructs were used in this SAXS analysis. SAXS data were first collected for a His-tagged variant of FN$^{\text{III}}$(1-2), His-FN$^{\text{III}}$(1-2). Results for this protein suggested that it was necessary to collect SAXS data on an un-tagged variant, FN$^{\text{III}}$(1-2), in order to confirm the modeling results (see below). The sequences of these variants are identical, with the exception of N-termini (Table 1).

Table 5: Sequences of the two FN$^{\text{III}}$(1-2) variants used in this study. The artificial sequences are highlighted in blue.

Variant	Sequence
His-FN$^{\text{III}}$(1-2)	GSSHHHHHHSSGLVPRGSHMSGPVEVFITETPSQPNSHPIQWNAPQPSHISKYILRWRPKNSVGRWKEATIPGHLNSYTIKGLKPGVVYEGQLISIQQYGHQEVTRFDFTTTSTSTPVTSNTVTGETTPFSPLVATSESVTEITASSFVVSWVSASDTVSGFRVEYELSEEGDEPQYLDLPSTATSVNIPDLLPGRKYIVNVYQISEDGEQSLILSTSQTTAPDA
FN$^{\text{III}}$(1-2)	GPHMSGPVEVFITETPSQPNSHPIQWNAPQPSHISKYILRWRPKNSVGRWKEATIPGHLNSYTIKGLKPGVVYEGQLISIQQYGHQEVTRFDFTTTSTSTPVTSNTVTGETTPFSPLVATSESVTEITASSFVVSWVSASDTVSGFRVEYELSEEGDEPQYLDLPSTATSVNIPDLLPGRKYIVNVYQISEDGEQSLILSTSQTTAPDA

5.2 The statistical conformations of His-FN$^{\text{III}}$(1-2)

5.2.1 His-FN$^{\text{III}}$(1-2) explores a larger conformational space as the ionic strength of the solution increases.

5.2.1.1 The His-FNIII(1-2) Guinier Rg increases with increasing ionic strength

The normalized scattering profiles for FN$^{\text{III}}$(1-2) at 0M NaCl, 0.15 M NaCl, and 0.50 M NaCl were markedly different (Figure 32A). A Guinier analysis was performed to determine the radius of gyration (R_g) of the protein in different solvent conditions. In a

Guinier analysis, the slope of an algebraic transformation of the data ((ln(I) vs. q^2) is directly proportional to the R_g (Svergun et al., 1987). The region of the scattering curve where a Guinier analysis is performed is dependent on the globularity of the molecule. I determined the R_g of His-FN^{III}(1-2) in 0M NaCl in the region $qR_g < 1.3$ (the reported Guinier region for globular proteins (Putnam et al., 2007)), and the R_g of His-FN^{III}(1-2) in 015M and 0.50M NaCl in the region $qR_g < 1.0$ (the reported Guinier region for spherically asymmetric and flexible proteins (Jacques et al., 2012)). The R_g, for His-FN^{III}(1-2) in 0M NaCl was 25.1 ± 0.37, in 0.15M NaCl it was 28.1 ± 0.90, and in 0.50M NaCl the R_g was 30.3 ± 0.91 (Figure 32B-D).

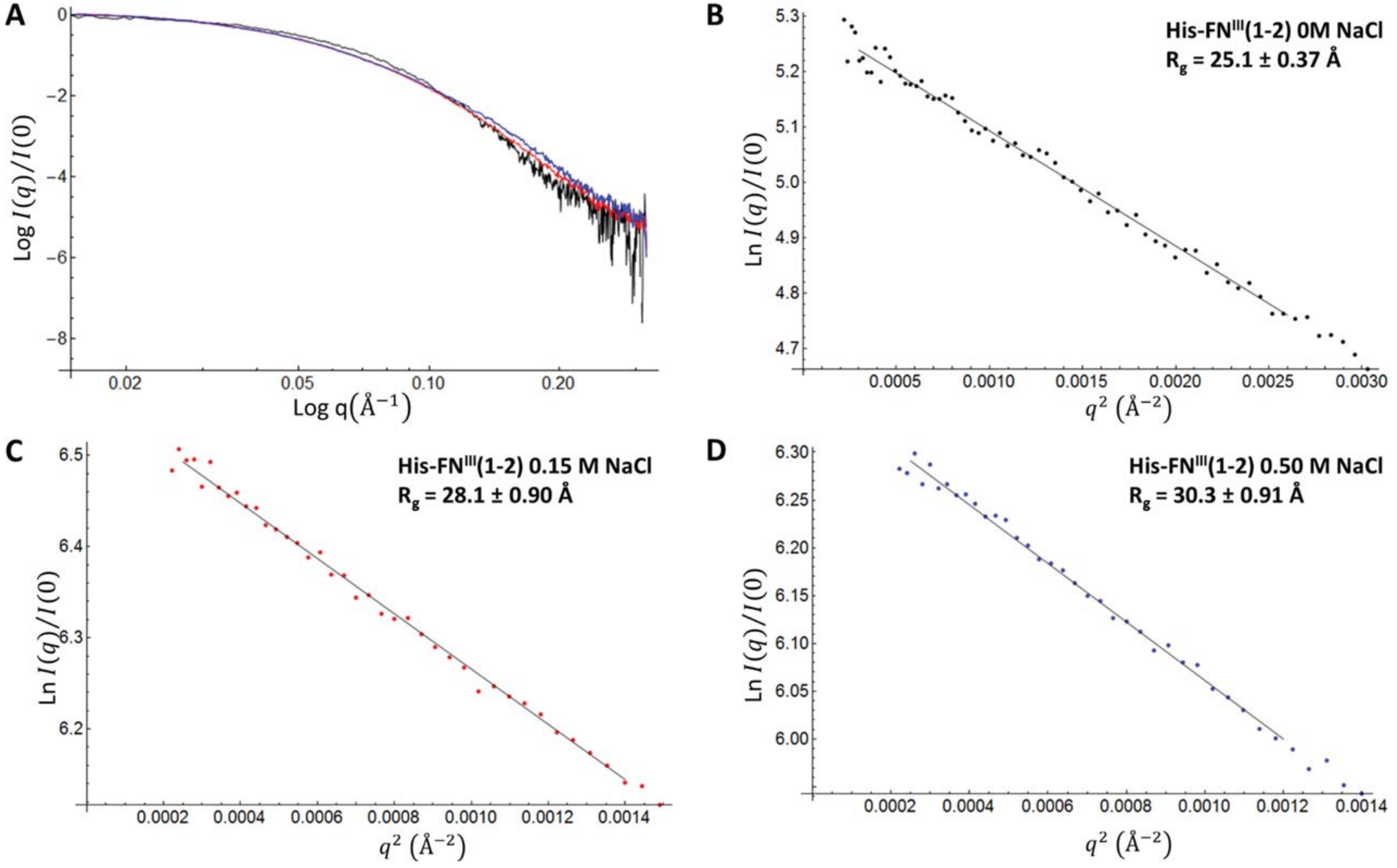
A
0
−2
−4
−6
−8
Log I(q)/I(0)
0.02
0.05
0.10
0.20
Log q(Å−1)
B
His-FNIII(1-2) 0M NaCl
Rg = 25.1 ± 0.37 Å
5.3
5.2
5.1
5.0
4.9
4.8
4.7
Ln I(q)/I(0)
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
q2 (Å−2)
C
His-FNIII(1-2) 0.15 M NaCl
Rg = 28.1 ± 0.90 Å
6.5
6.4
6.3
6.2
Ln I(q)/I(0)
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
q2 (Å−2)
D
His-FNIII(1-2) 0.50 M NaCl
Rg = 30.3 ± 0.91 Å
6.30
6.25
6.20
6.15
6.10
6.05
6.00
Ln I(q)/I(0)
0.0002
0.0004
0.0006
0.0008
0.0010
0.0012
0.0014
q2 (Å−2)

Figure 32: SAXS profiles and Guinier analysis of His-FN^{III}(1-2) in three different ionic strength conditions. A) The SAXS profiles for His-FN^{III}(1-2) in 0M NaCl (black), 0.15M NaCl (red), and 0.50M NaCl (blue). B-D) Guinier analysis of His-FN^{III}(1-2). Solid line is the best fit line to the data. B) His-FN^{III}(1-2) in 0M NaCl, C) His-FN^{III}(1-2) in 0.15M NaCl, D) His-FN^{III}(1-2) in 0.50M NaCl. Excellent linear correlations are observed within the Guinier region for all three datasets, indicating that each sample was free of self-association or inter-particle interference.

The Guinier analysis indicated demonstrated that the conformational space explored by His-FN^{III}(1-2) increases at higher ionic strength, as evidenced by the increasing Guinier R_g (Figure 32B-D). It should be noted that the CRYSOL-calculated R_g for the NMR model (pdb: 2HA1.1) was 22.4 Å. This is lower than any of the experimentally calculated R_g's. This discrepancy is probably the result of the addition of a 20-residue N-terminal His tag in His-FN^{III}(1-2). The change in R_g as a function of ionic strength is intriguing. It indicates that the overall dimensions of the statistical conformation increase as a function of ionic strength.

5.2.1.2 There may be flexibility in both the inter-domain linker and the N-terminal His-tag.

In order to assess the globularity of His-FN^{III}(1-2), I performed a Kratky and Porod-Debye analysis of the scattering data. These results are presented in Figure 33.

The dimensionless Kratky and Porod-Debye plots are insufficient to assess the globularity of His-FN^{III}(1-2) in any of the three solution conditions. Recall from Chapter 2 that in order to be defined as globular, a molecule must both exhibit a peak in the dimensionless Kratky plot at $\sqrt{3}$ and have a plateau in the Porod-Debye plot. None of the SAXS data of His-FN^{III}(1-2) meet both these requirements. These results strongly suggest that at least a portion of His-FN^{III}(1-2) is flexible or spherically asymmetric in all solution conditions. However, it is unknown whether this suggested spherical asymmetry or flexibility is the result of flexibility in the N-terminal His-tag, interdomain flexibility, or both.

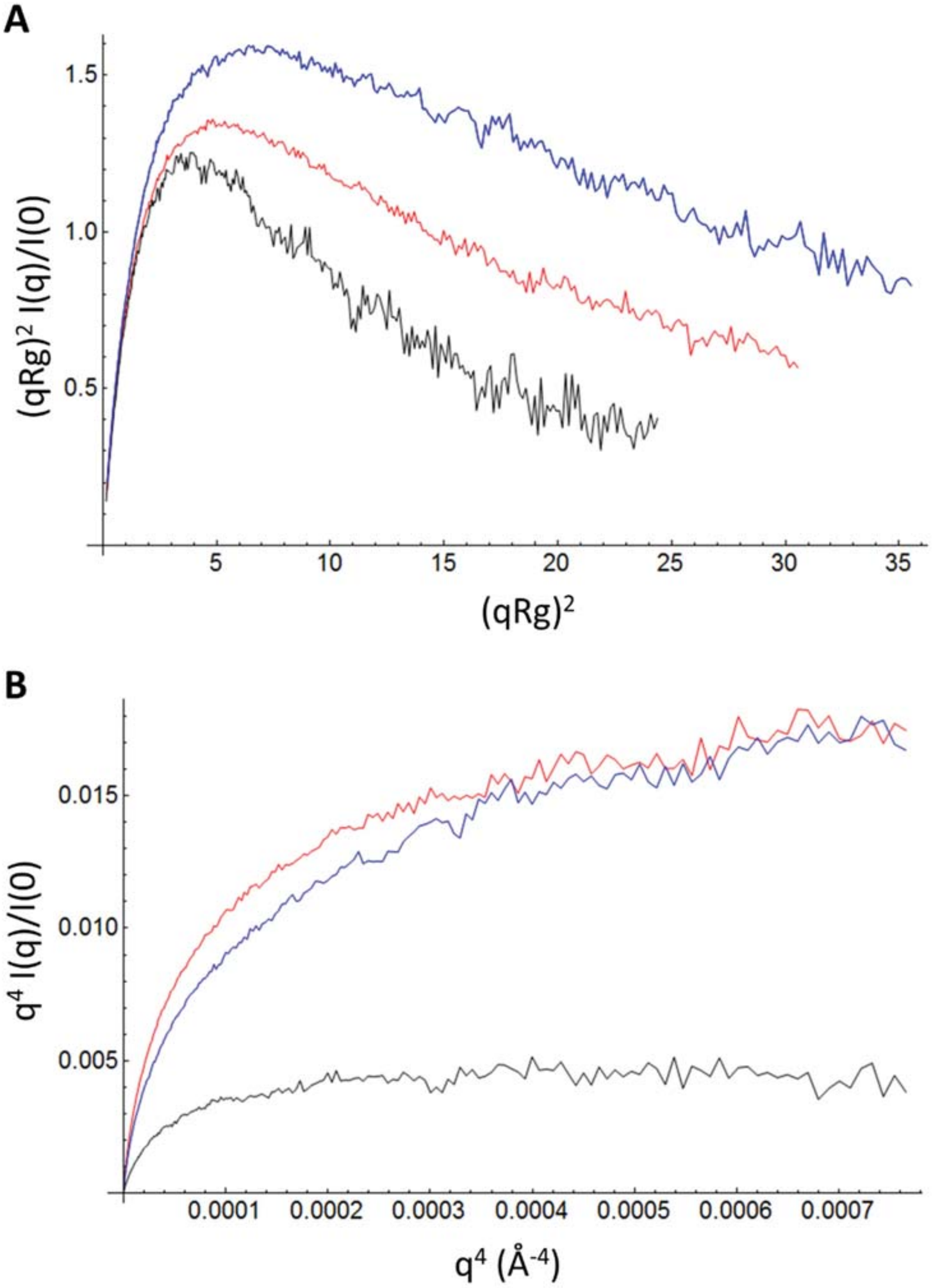

Figure 33: Kratky and Porod-Debye analysis to assess globularity of His-FN^{III}(1-2) in different solution conditions: His-FN^{III}(1-2) in 0M NaCl (black), in 0.15M NaCl (red), and in 0.50M NaCl (blue). A) The dimensionless Kratky plot does not

exhibit a peak at $\sqrt{3}$ for any of the SAXS data. This result indicates that His-FN^{III}(1-2) may be non-globular in these conditions. B) The Porod-Debye plot for His-FN^{III}(1-2) in 0M NaCl (black) exhibits a hyperbolic plateau that can be indicative of globularity. Neither His-FN^{III}(1-2) in 0.15M NaCl (red) or His-FN^{III}(1-2) in 0.50M NaCl (blue) exhibit a plateau, indicating that His-FN^{III}(1-2) may be non-globular in these conditions.

5.2.2 His-FN^{III}(1-2) in 0M NaCl is best described by a static polymer model, while His-FN^{III}(1-2) in 0.15M and 0.50M NaCl are best described as flexible polymers

Next, I fit the experimental SAXS data to a variety of polymer models in order to describe the low-resolution shape of the molecule. This analysis is helpful in determining if a molecule is best described by a uniformly dense static model or a spherically asymmetric model. The best fit polymer model to the scattering data of His-FN^{III}(1-2) in 0M NaCl is an ellipsoid model (Figure 34) (Svergun et al., 1987). This model has two parameters: the radius of the long axis (35.4 ± 0.19Å) and the radius of the short axis (11.3 ± 0.19Å). The χ^2 statistic for this fit is 1.83, indicating that the model is a reasonable fit of the data. The model fit the data poorly above $q > 0.12$ Å (Figure 34A, bottom). When the NMR model (pdb: 2HA1.1) is aligned with an ellipsoid of this size, neither the mass of the inter-domain linker or the N-terminal His-tag is enclosed by the ellipsoid (Figure 4). Additionally, the largest dimension of a single fibronectin domain, determined from the NMR model is 43.6 Å, corresponding to a q-value of 0.144 Å^{-1}. At $q \geq 0.144$ Å^{-1} intra-domain scattering dominates the scattering profile and a static polymer model is not sensitive to this intra-domain scattering. The poor fit of the data to the model at $q > 0.12$ Å could be due to both intra-domain scattering and the specific conformations of the inter-domain linker or N-terminal His-tag. Nevertheless, the fit of the ellipsoid model to the data strongly suggests that the SAXS data of His-FN$^{\mathrm{III}}$(1-2) in 0M NaCl is consistent with the NMR model, where the two fibronectin domains contact each other.

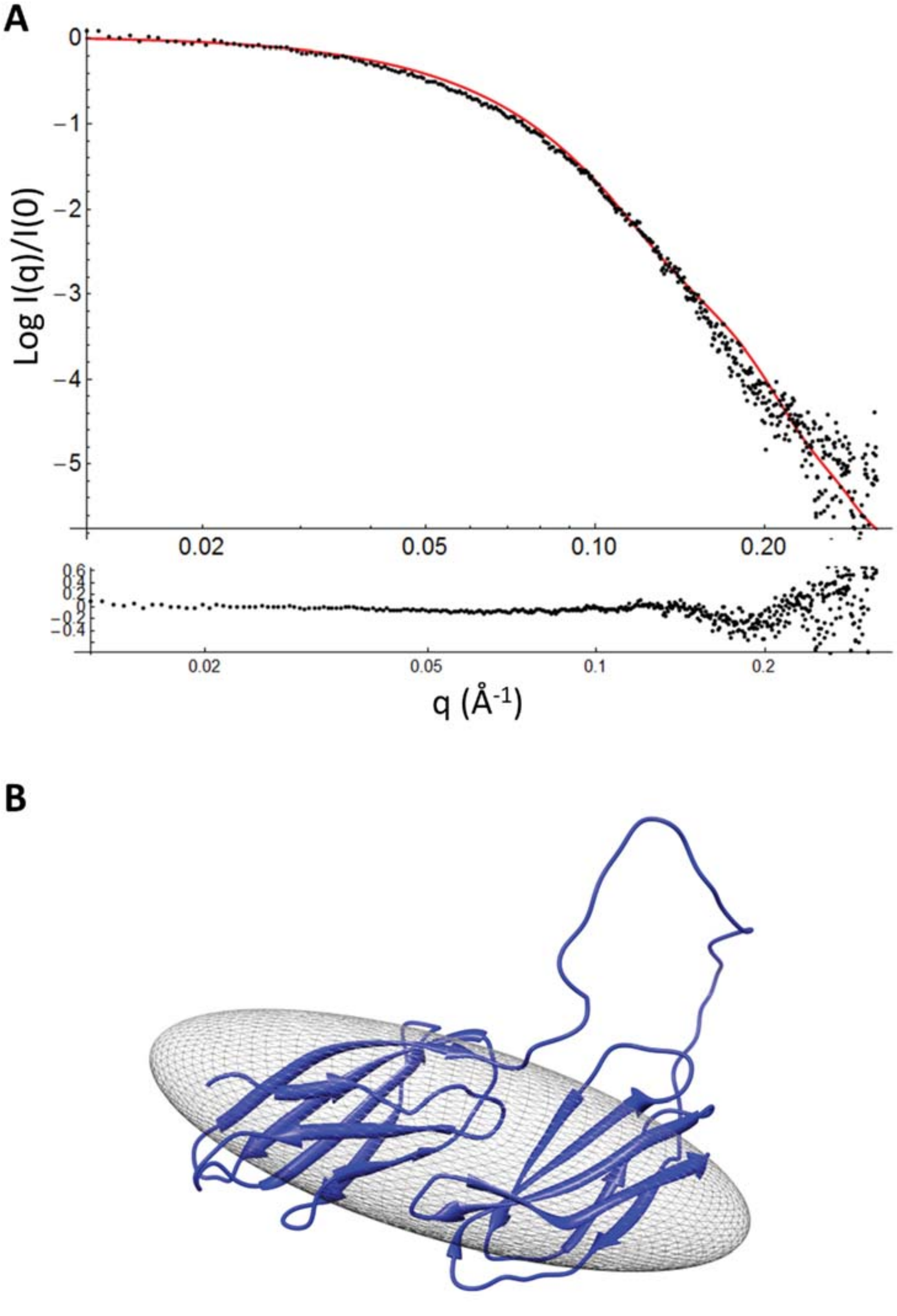

Figure 34: Fit of an ellipsoid model (red) to the His-FN^{III}(1-2) in 0M NaCl SAXS data (black). A) Top: fit of the ellipsoid model (red) to the SAXS data (black). $X^2 = 1.83$.

B) The NMR model of FN^{III}(1-2) superimposed on an ellipsoid of dimensions 35.4 x 11.3 Å.

In contrast, a static uniformly dense model was not a good fit of the His-FN^{III}(1-2) SAXS data in 0.15M and 0.5M NaCl. Instead, the scattering function of barbell model best fit the SAXS data from these two conditions. The barbell model and its' scattering function are described in Chapter 4.5. Essentially, the model describes scattering from two identical spheres held at a fixed distance from one another. The linker between the two spheres is modeled as an excluded volume Gaussian coil and the mass of the linker is explicitly accounted for in the scattering function. The fitted-for parameters in this model are the radius of the spheres and the center-to-center distance between spheres.

The best-fit parameters for the His-FN^{III}(1-2) 0.15M NaCl SAXS data are a radius of 20.6 ±0.03Å for the spherical domains, and a center-to-center distance of 66.0 Å, with $\chi^2 = 1.05$. The model does not account of all of the features observed in the experimental scattering profile (Figure 35A). However, based on the goodness-of-fit statistic, it is still a reasonable model of the data. The difference between the models scattering profile and the data at high scattering angles could be accounted for by the non-sphericty of the domains. The longest dimension of a fibronectin domain is 43.6 Å, and at scattering angles between $0.15Å^{-1} < q < 0.3 Å^{-1}$ (corresponding to real-space dimensions of 21-42Å) intra-domain scattering is likely to dominate. The calculated R_g of this model, using a 66Å center-to-center distance between spheres, is 33 Å. However, only the mass of the spherical domains are considered in this R_g function and not the mass of the inter-domain linker or N-terminus is (Kaya, 2004). The R_g calculated from the barbell model is larger than the calculated Guinier 28 Å. The difference in R_g could be due to the mass and arrangement of the inter-domain linker and the N-terminus. Although the fit of the barbell model to the data is not perfect, this spherically asymmetric model still fit the experimental scattering data better than a uniformly dense static polymer model. The modeling results suggest that His-FN^{III}(1-2) predominately populates an expanded conformation in 0.15M NaCl. The two domains are separated by an average distance of approximately 25 Å.

Likewise, the scattering data for His-FN^{III}(1-2) 0.50M NaCl was best fit by the barbell model (Figure 35B). The best fit parameters for this model are 20.3 ± 0.03 Å for the spherical domains, and 71.1 ± 0.31 Å for the center-to-center distance between domains. The χ^2 statistic is 1.01. The radius of the spherical domains is 0.3 Å smaller than the radius determined from the 0.15M NaCl dataset. Because these two datasets were fit independently, the nearly-identical domain dimensions support the use of this model to fit the data. The center-to-center distance between spherical domains is 71.1 ± 0.3 Å which corresponds to an R_g of 35.5 Å. Again, this is larger than the Guinier R_g, but the difference could be accounted for by the additional mass distribution of the inter-domain linker and N-terminus that are not accounted for in the R_g calculated using the barbell model. This fit of the barbell model to the data suggests that at 0.50M NaCl the two His-FN^{III}(1-2) domains do not fully populate a conformation with a desolvated interface.

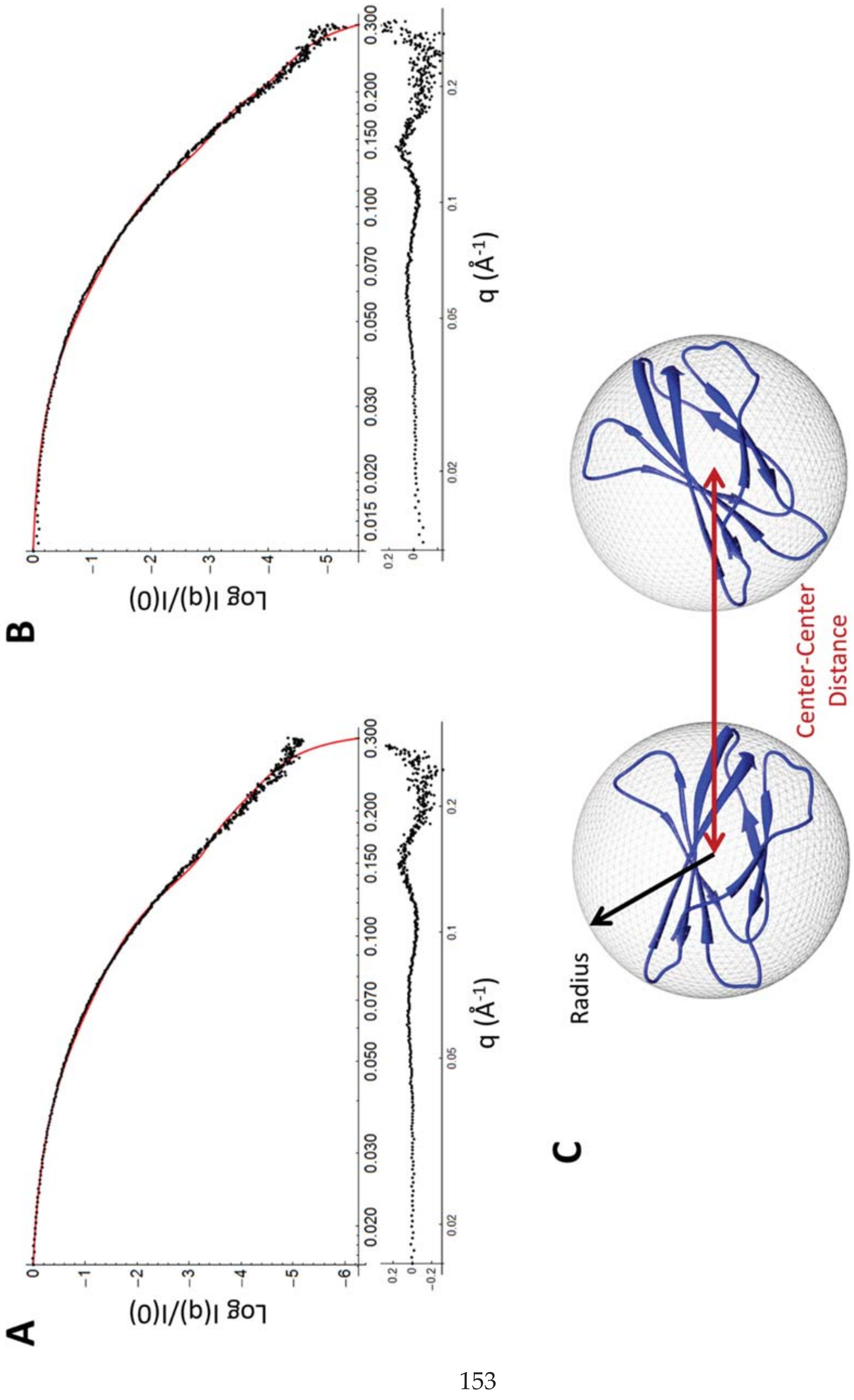
A
B
C
Log I(q)/I(0)
0
-1
-2
-3
-4
-5
-6
0.015
0.020
0.030
0.050
0.070
0.100
0.150
0.200
0.300
0.2
0
-0.2
0.02
0.05
0.1
0.2
q (Å-1)
Radius
Center-Center
Distance

Figure 35: Fit of the barbell model (red) to the His-FN^{III}(1-2) 0.15M and 0.50M NaCl data (black). A) Fit of His-FN^{III}(1-2) 0.15M NaCl SAXS data to the barbell model. $\chi^2 = 1.05$. B) Fit of the His-FN^{III}(1-2) 0.50M NaCl SAXS data to the barbell model. $\chi^2 = 1.01$. C) Fibronectin domains (pdb: 2HA1) superimposed on a sphere of 20 Å. The parameters for the barbell model are indicated. The barbell model adequately describes the scattering in 0.15M and 0.50M NaCl. These results suggest that, on average, His-FN^{III}(1-2) does not form a desolvated interface.

The results of the polymer modeling suggest that His-FN^{III}(1-2) in 0M NaCl is best described by a single static model, while at higher ionic strength the protein explores a larger region of conformational space and, on average, the two domains are not in permanent contact with each other. The polymer modeling is consistent with the Guinier analysis above. It is also consistent with the Erickson lab data that demonstrates that at physiological ionic strength FN^{III}(1-2) does not fully populate a compact conformation consistent with the NMR model. In order to further verify this conclusion, a theoretical SAXS profile calculated from the NMR model was compared to the experimental data.

5.2.3 An atomistic rigid body model fits the His-FN^{III}(1-2) 0M NaCl data, but not the SAXS data of His-FN^{III}(1-2) in higher ionic strength conditions.

CRYSOL was used to calculate a theoretical aggregate SAXS profile from the 2HA1 structural ensemble (Svergun et al., 1995). The fit of the profile and the experimental data were optimized using adjustable parameters for the hydration layer (see Chapter 2.3). The fit of the model to the data is presented in Figure 36A. The

CRYSOL χ^2 is 0.82, and the residuals are nearly stochastic in the entire q-range. The fit of the model to the data indicates that the NMR model is a reasonable model of the His-FN^{III}(1-2) statistical conformation in 0M NaCl, and at this ionic strength, FN^{III}(1-2) predominately populates a compact conformation.

The theoretical aggregate SAXS profile of the NMR model was also compared to the experimental data in 0.15M and 0.50M NaCl. The results are presented in Figure 36B and C. The CRYSOL χ^2 of the fit of the 0.15 M NaCl data to the NMR model is 2.25, and the CRYSOL χ^2 for the fit of the 0.5M NaCl data to the NMR model is 2.51. In both instances, the overall dimensions (visible in the low-q region of the scattering curve) are smaller for the theoretical SAXS curve calculated from the NMR model than they are in the SAXS data. The rigid body analysis confirms that the His-FN^{III}(1-2) NMR model is inconsistent with the SAXS data in higher ionic strength conditions, and suggests expansion of the statistical conformation at higher ionic strengths.

The published NMR data was collected at pH 7.0 with 0.15M NaCl. It should be noted that the Erickson lab biophysical experiments were performed in phosphate-buffered saline, pH 7.4, as were the SAXS experiments. The His-FN^{III}(1-2) SAXS data supports the Erickson lab conclusions that the two domains do not predominately populate a compact conformation in 0.15M NaCl. The pH difference in the experimental conditions could result in a different statistical conformation, though this is unlikely. It is possible that the NOE constraints that led to an NMR structure with a static interface between domains were actually the result of transient interactions between domains that are not held by a rigid linker. Indeed, this hypothesis is supported by the NMR data (Vakonakis et al., 2007). The chemical shift perturbations caused by FN^{III}1 and FN^{III}2 domains were small, and the initial NOE and residual dipolar coupling results were ambiguous. Two distinct conformations fit the data equally well. The interaction surface in FN^{III}1 was different in both models. These results suggest that the NOE's observed in the experiment could be the result of transient domain:domain interactions.

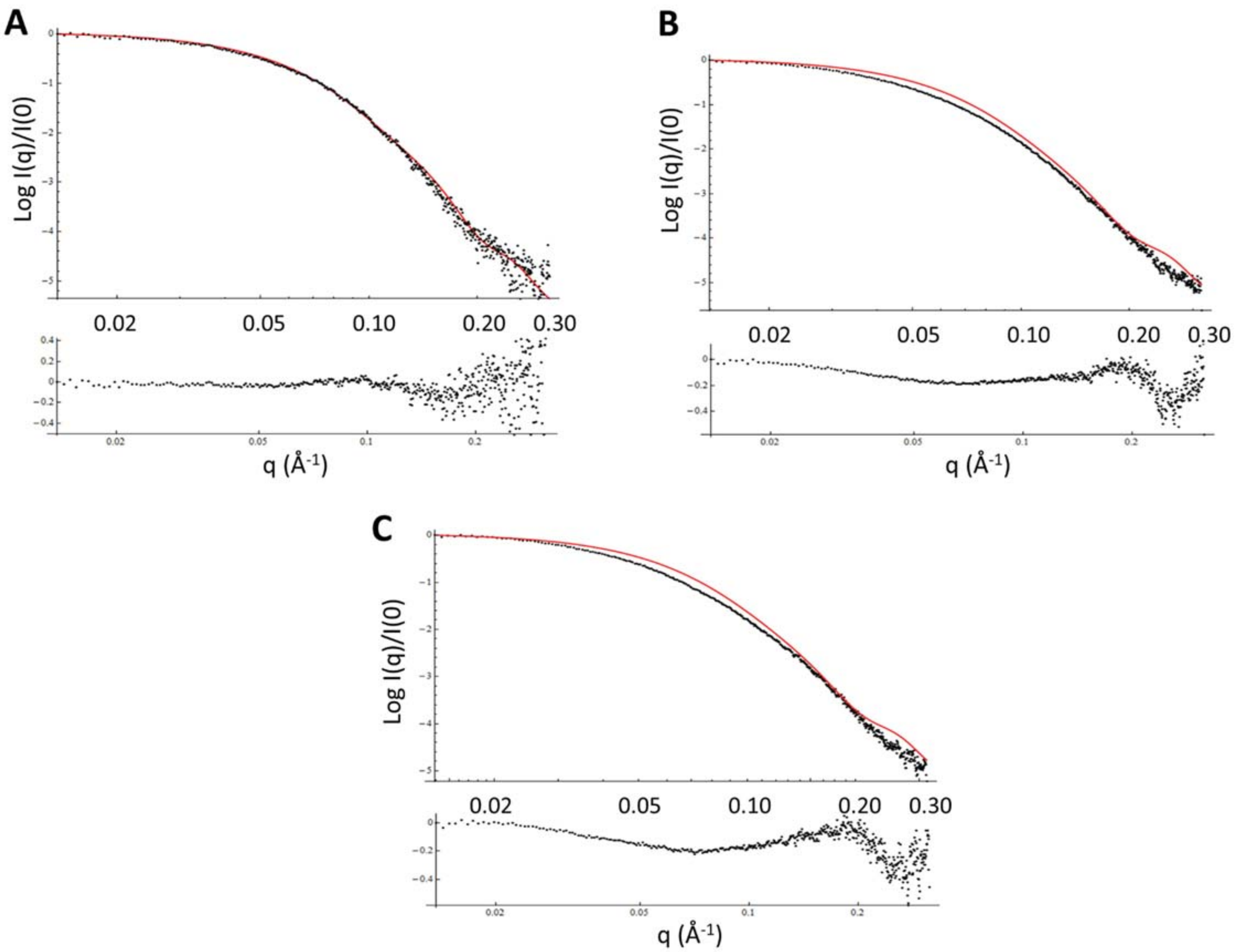
A
B
C
Log I(q)/I(0)
0
-1
-2
-3
-4
-5
0.02
0.05
0.10
0.20
0.30
0.4
0.2
0
-0.2
-0.4
0.02
0.05
0.1
0.2
q (Å^{-1})

Figure 36: Comparison of the His-FN^{III}(1-2) SAXS data (black) to a theoretical scattering profile calculated from the NMR structural ensemble (red). A) Fit of the theoretical scattering curve to His-FN^{III}(1-2) in 0M NaCl. $\chi^2 = 0.82$. B) Fit of the theoretical scattering curve to His-FN^{III}(1-2) in 0.15M NaCl. $\chi^2 = 2.25$. C) Fit of the theoretical scattering curve to His-FN^{III}(1-2) in 0.50M NaCl. $\chi^2 = 2.51$. The theoretical scattering profile of the NMR model can describe the scattering data at 0M NaCl, but the statistical conformation is expanded in 0.15M or 0.50M NaCl.

5.2.4 Ensemble analysis of His-FN^{III}(1-2) in 0.15M and 0.50M NaCl

5.2.4.1 The aggregate scattering profile from a RanCH-generated ensemble is a poor fit to the experimental data.

The polymer physics modeling of the His-FN^{III}(1-2) SAXS data in 0.15M and 0.50M NaCl indicated that the statistical conformations can be described by a barbell polymer model. This model provides an average center-to-center distance between domains. However, it does not provide any insights into the statistical conformation other than the average center-to-center distance between domains and the spherically averaged size of the domains. Ensemble modeling may provide distributional information about the statistical conformation not provided by the polymer model. Specifically, ensemble modeling could provide additional distributional information about R_g and D_{max}.

RanCH (Bernado et al., 2007)was used to generate a flexible 10,000 member ensemble in which the domain orientations, inter-domain linker, and N-terminus were randomized. The aggregate scattering profile from the ensemble is compared to the experimental scattering curves in Figure 37. The χ^2 for the fit of the aggregate scattering profile to the 0.15M NaCl dataset is 1.13, and the χ^2 for the fit of the 0.50M NaCl dataset is 1.08. In both cases, the intensity of the aggregate scattering profile is greater than that for the data at q > 0.10 Å^{-1}. This "flattening" of the scattering profile suggests that the statistical conformation of the RanCH-generated ensemble is more elongated than suggested by the data. In other words, the fit of the aggregate scattering curve to the data indicates that the statistical conformation is more compact than would be expected if only hard sphere constraints limit the conformational space of His-FN^{III}(1-2). These results indicate that enumeration of a minimal ensemble from the data will yield additional information about the statistical conformation (see Section 2.5.4 in Chapter 2).

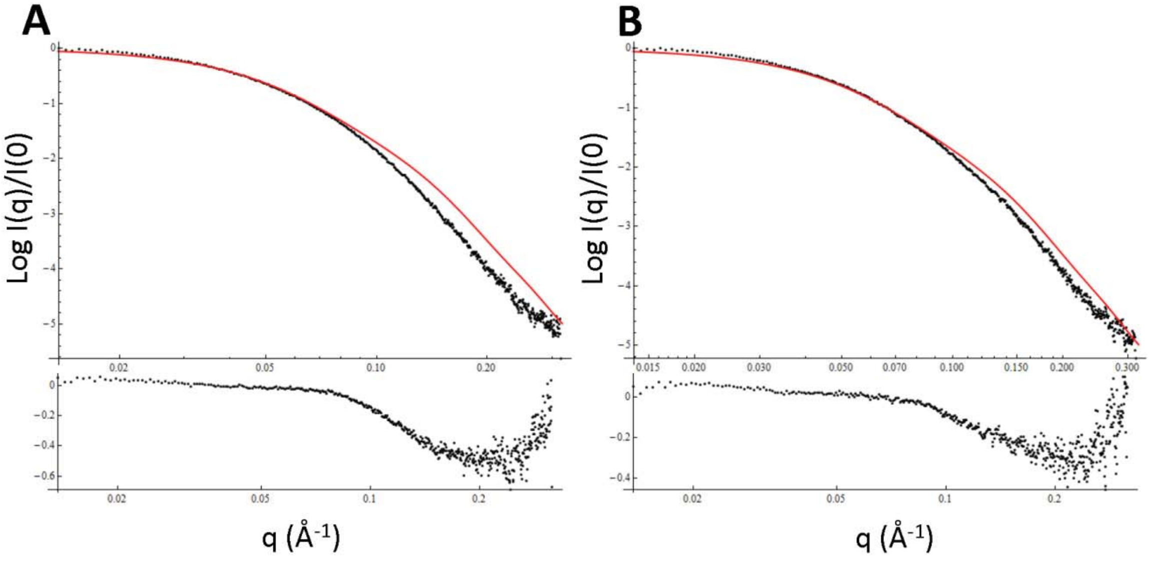

Figure 37: Fit of the aggregate scattering profile from the RanCH generated ensemble (red) to the SAXS data (black). A) Fit of the aggregate scattering profile to the 0.15M NaCl data.χ^2 = 1.13 B) Fit of the aggregate scattering profile to the 0.50M

NaCl data. $\chi^2 = 1.13$. Because the aggregate scattering profile does not fully describe the data, enumeration of a minimal ensemble is warranted.

5.2.4.2 The R_g and D_{max} distributions of the minimal ensembles confirm that His-FN^{III}(1-2) explores a larger region of conformational space as ionic strength increases.

GAJOE was used to select a minimal ensemble consistent with the SAXS data. The results are presented in Figure 38. The selection of the minimal ensembles was robust: three independent calculations of a minimal ensemble resulted in similar R_g and D_{max} distributions. The calculated minimal ensembles for the SAXS profile of His-FN^{III}(1-2) in 0.15M NaCl had an average R_g of 27. 15 Å with a range between 18.5 – 44 Å for all ensemble members. This is considerably smaller than the R_g distribution of 10,000 member parent ensemble (17.25 – 54 Å) indicating that there are favorable protein:protein interactions within the His-FN^{III}(1-2) statistical conformation in concentrations of 0.15M NaCl. The D_{max} distributions of the minimal ensembles indicated similar compactness (Figure 38E). It must be noted, however, that the minimal ensemble results ensemble are not consistent with the NMR model. There is a range of conformations in solution that, on average, are more elongated than the conformations in the NMR data derived model.

Similarly, the minimal ensembles derived from the RanCH parent ensemble fit the His-FN^{III}(1-2) 0.50M SAXS data well. The average R_g of the best-fit minimal ensemble was 29.85 Å and the R_g distribution was between 19.5 and 44 Å (Figure 38D). This distribution is very similar to the best-fit distribution of the minimal ensemble constrained against the His-FN^{III}(1-2) 0.15M SAXS data. However, the mean, variance, and skewness of the two R_g distributions are not identical (Compare Figures 38C and D). The statistical conformation of His-FN^{III}(1-2) in 0.50M NaCl explores a larger breadth of conformational space that the 0.15M NaCl statistical conformation. When the D_{max} distributions are compared, we also see that this is the case (Figures 38E and F). Additionally, the R_g and D_{max} distributions of the minimal ensembles for the His-FN^{III}(1-2) 0.50MNaCl fit are only slightly smaller than that of the entire 10,000 member parent ensemble. These results are consistent with the polymer physics modeling results. Additionally, ensemble modeling resulted in additional information about the statistical conformation that was not accessible by polymer physics modeling: the R_g and D_{max} distributions. This is an example where judicious use of ensemble modeling is appropriate.

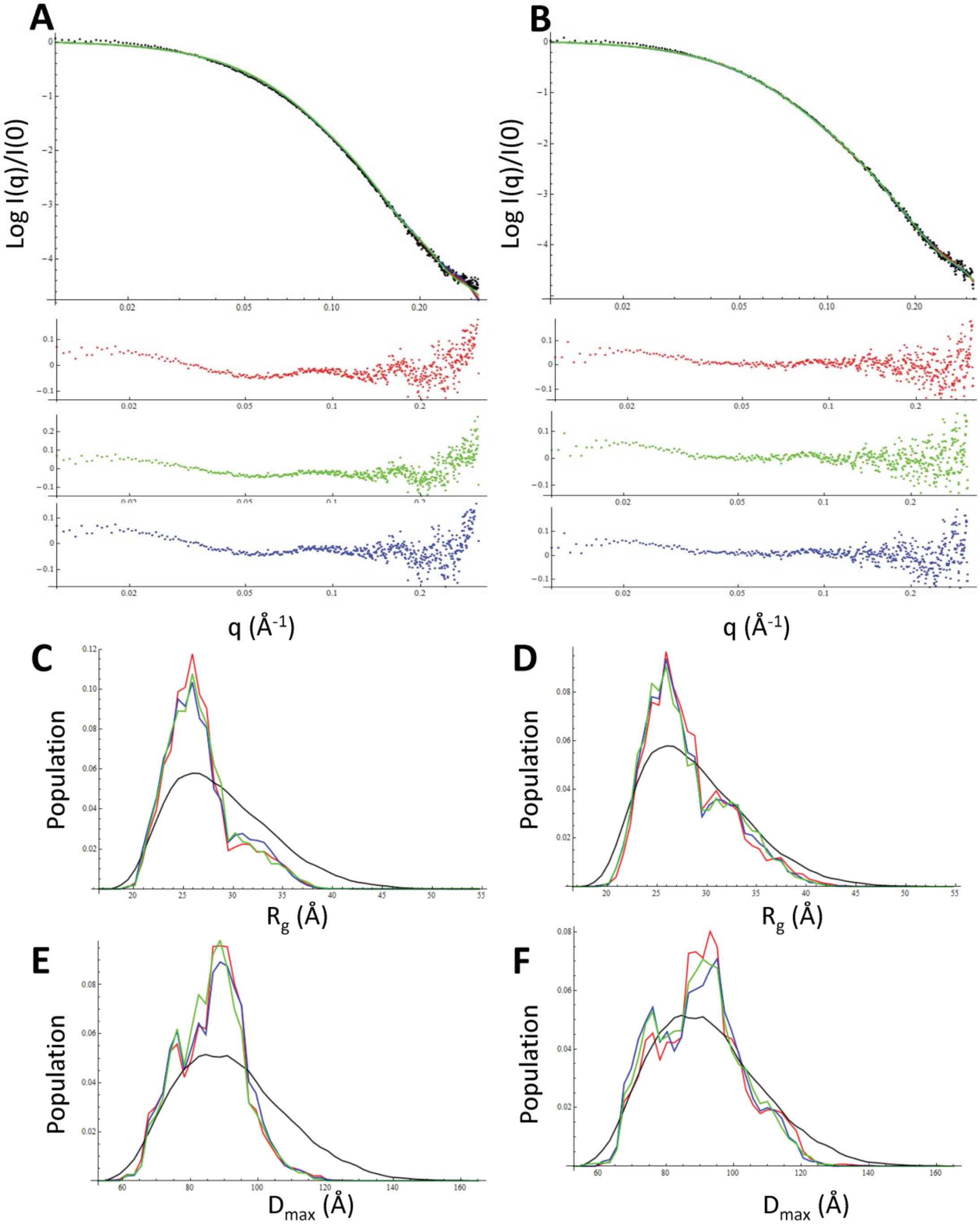
A
B
Log I(q)/I(0)
q (Å^{-1})
C
D
Population
R$_g$ (Å)
E
F
D$_{max}$ (Å)
0.02
0.05
0.10
0.20
20
25
30
35
40
45
50
55
60
80
100
120
140
160

Figure 38: Minimal ensemble analysis of His-FN^{III}(1-2) in 0.15M and 0.5M NaCl. A) Fit of the aggregate minimal SAXS profile from the minimal ensembles (red, green, blue) to the 0.15M SAXS data (black). Residuals are below the plot. B) Fit of the aggregate minimal SAXS profile from the minimal ensembles (red, green, blue) to the 0.50M SAXS data (black). Residuals are below the plot. C) The R_g distribution of the best fit minimal ensembles (red, green, blue) to the 0.15M SAXS data and the parent ensemble (black). D) The R_g distribution of the best fit minimal ensembles (red, green, blue) to the 0.50M SAXS data and the parent ensemble (black). E) The D_{max} distribution of the best fit minimal ensembles (red, green, blue) to the 0.15M SAXS data and the parent ensemble (black). F) The D_{max} distribution of the best fit minimal ensembles (red, green, blue) to the 0.50M SAXS data and the parent ensemble (black). The ensemble modeling results indicate that at both the dimensions of the statistical conformation increase as the ionic strength increases.

5.3 FN^{III}(1-2) SAXS data confirms the conclusions obtained by the analysis of the His-FN^{III}(1-2) SAXS data

To verify that the results obtained in section 5.2 did not reflect expansion of the N-terminal His-tag alone, a I performed a complete SAXS analysis of a non-His-tagged variant of FN^{III}(1-2) in 0.15M and 0.50M NaCl. The SAXS data is noisy and relatively featureless (Figure 39A and B) due to the low flux of the SAXS instrument and the protein concentration of the sample, so only the R_g of the the statistical conformation can be assessed using this data. The R_g obtained from a Guinier analysis of the data for FN^{III}(1-2) in 0.15M NaCl was 25.9 ± 0.004Å, and the R_g of FN^{III}(1-2) in 0.50M NaCl was 27.8 ± 0.004Å. These R_g values are ~ 2Å lower than those of His-FN^{III}(1-2) under the same conditions, which would be expected given the truncation of the flexible N-terminal tag. The R_g of FN^{III}(1-2) increases by 2Å when the ionic strength of the solution is increased from 0.15M to 0.50M NaCl. These results are compared to the R_gs' of His-FN^{III}(1-2) in 0.15M and 0.50M NaCl in Table 6, below. There is a similar increase in both variants as the ionic strength is increased. This result is consistent with FRET experiments that showed there was a reduction in FRET efficiency as the NaCl concentration of the solution was increased 0.20M to 0.50M (Ohashi and Erickson, 2011).

Table 6: Rg comparison of His-FN^{III}(1-2) and FN^{III}(1-2)

Fibronectin Variant	**R_g in 0.15M NaCl**	**Rg in 0.50M NaCl**
His-FN^{III}(1-2)	28.1 ± 0.9 Å	30.3 ± 0.9 Å
FN^{III}(1-2)	25.9 ± 0.04 Å	27.8 ± 0.04 Å

Presented in Figure 39 is a summary of the results from the last stage of the SAXS analysis of FN^{III}(1-2): enumeration of a minimal ensemble consistent with the SAXS data. Interestingly, there is very little difference in the R_g and D_{max} distributions that best fit the 0.15M and 0.50M SAXS data (Figure 39C-E). The average R_g of the 0.15M minimal ensemble is 27.3Å, and the average R_g of the 0.50M minimal ensemble is 28.4Å. The difference in the average R_g calculated by the Guinier analysis and that calculated from the minimal ensembles can be explained by the relatively noisy and featureless SAXS data. It is possible that many different ensembles with a wide range of R_gs can be fit equally well to the SAXS data. Additional data is needed to verify the FN^{III}(1-2) SAXS ensemble modeling results. Nevertheless, the breadth of the R_g and D_{max} distributions, as well as the average $R_{g,}$ in both solution conditions suggests that, on average, $FN^{III}(1\text{-}2)$ populates an expanded conformation.

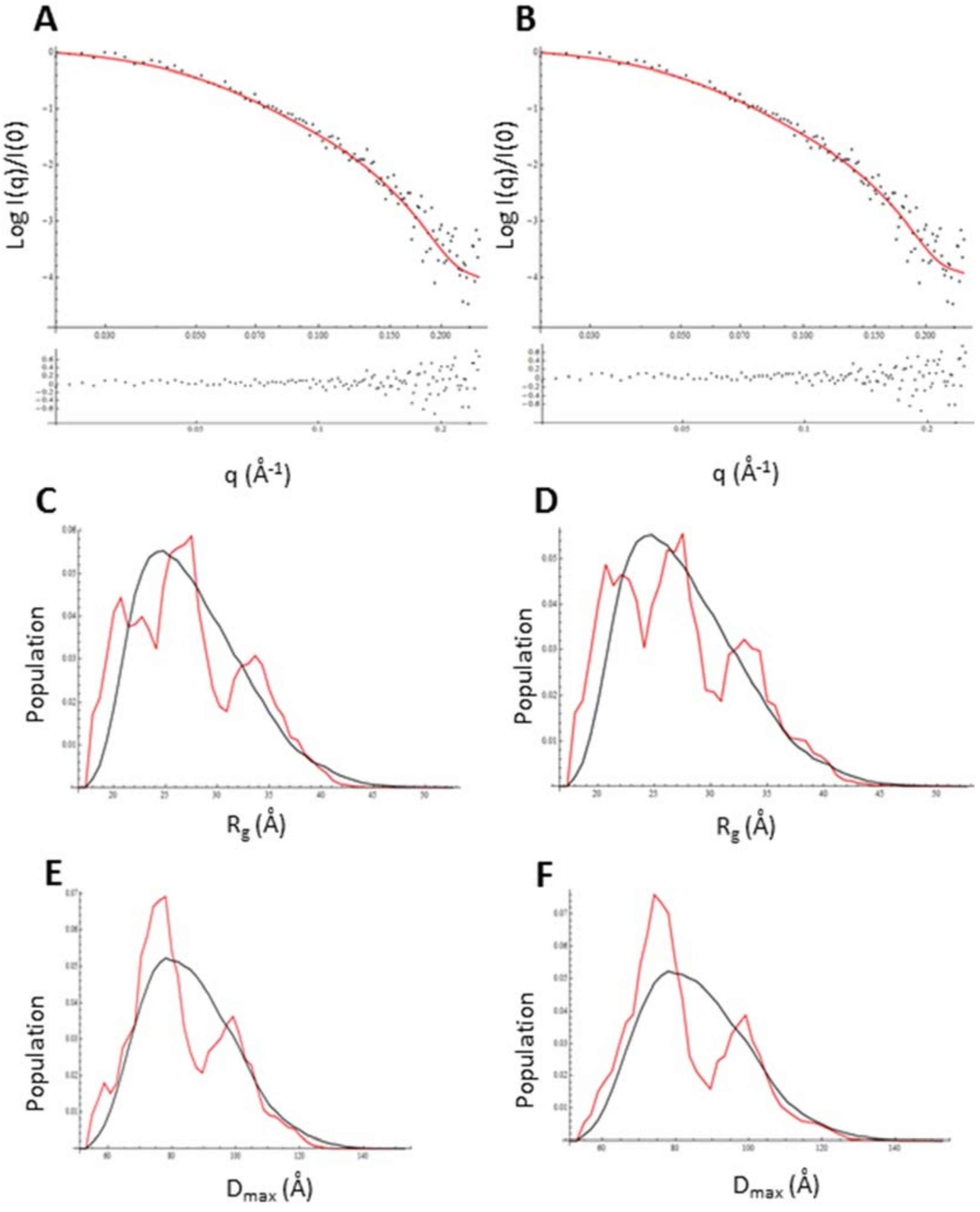

Figure 39: Minimal ensemble analysis of FN^{III}(1-2) in 0.15M and 0.5M NaCl. A) Fit of the aggregate minimal SAXS profile from the minimal ensemble (red) to the 0.15M SAXS data (black). Residuals are below the plot. B) Fit of the aggregate minimal SAXS profile from the minimal ensembles (red) to the 0.50M SAXS data

(black). Residuals are below the plot. C) The R_g distribution of the best fit minimal ensemble (red) to the 0.15M SAXS data and the parent ensemble (black). D) The R_g distribution of the best fit minimal ensembles (red) to the 0.50M SAXS data and the parent ensemble (black). E) The D_{max} distribution of the best fit minimal ensembles (red) to the 0.15M SAXS data and the parent ensemble (black). F) The D_{max} distribution of the best fit minimal ensembles (red) to the 0.50M SAXS data and the parent ensemble (black). The ensemble modeling results indicate that at both 0.15M NaCl and 0.5M NaCl the conformational space of FN^{III}(1-2) is very large.

5.4 Discussion

The ensemble modeling results for both His-FN^{III}(1-2) 0.15M and 0.50M NaCl SAXS data and FN^{III}(1-2) 0.15M and 0.50M NaCl SAXS data, combined with the analysis in sections 5.2.1-5.2.3, demonstrate that when the ionic strength of the solution increases, the conformational space explored by the FN^{III}(1-2) statistical conformation also expands. The most compact statistical conformation is observed in 0M NaCl and the theoretical SAXS profile derived from the NMR structural ensemble is consistent with the SAXS data in 0M NaCl.

We must ask, what are the contributing properties of the protein and solution that result in this observed expansion of conformational space? Both solvation and charge-screening of the interface between the two domains and the 35 residue linker can contribute to expansion of conformational space.

Most likely, electrostatic interactions at the interface of FN^{III} domains 1 and 2 help to stabilize the interface at low ionic strengths. At pH 7.4 the overall net charge of domain 1 is positive (pI = 9.45) and the overall net charge of domain 2 is negative (pI =3.79). As the ionic strength of the solution increases the resultant ionic screening should destabilize the salt bridge that may be present at low ionic strength. Campbell and colleagues fit the initial NMR data to two different models: one in which a salt-bridge stabilizes the interactions between domains 1 and 2, and another model in which other non-specific charge interactions constrain and stabilize the compact conformation. The observed NOEs were small and the initial NOE and RDC data were ambiguous, resulting in two models that fit the data equally well. It is possible that the observed NOEs were the result of transient domain:domain interactions. By using the NOEs as distance constraints, the researchers may have been biasing the data toward compact conformations, when, in reality, the FN^{III}(1-2) statistical conformation may be composed of conformers lacking domain:domain interactions in addition to compact conformers.

The 0M NaCl SAXS results are consistent with the NMR data collected in 0.15M NaCl suggesting that at very low ionic strength, a compact region of conformational space is populated. However, once the ionic strength is increased to near-physiological ionic strength, the charged ions could be shielding surface charges of the domains, resulting in an expanded statistical conformation.

Charge shielding in the linker region could also be responsible for the observed expansion of the statistical conformation, particularly in higher ionic strength conditions. However, given the paucity of ionized groups in the linker, the impact of charge shielding is expected to be minimal. The sequence of the 35 residue linker is presented below in Table 7. The charged residues are highlighted red. The composition of the linker is polar and the fraction of charged residues (FCR) is 0.08, suggesting that the linker is most likely in a coil conformation instead of a collapsed globule (Das and Pappu, 2013). Therefore, the inter-domain linker is a weak polyampholyte. It is unlikely that an increase in ionic strength results in an expansion of the inter-domain linker.

Table 7: The sequence of the inter-domain linker

Sequence of the Interdomain Linker
T S T S T P V T S N T V T G E T T P F S P L V A T S E S V T E I T A S S

Thus, the expansion of the FN^{III}(1-2) as a function of ionic strength could be reflective of the relative populations of two distinct statistical conformations: one compact conformation where electrostatic interactions between domains constrain the conformational space of the protein and another set of expanded conformations that results from charge screening of the domains. Once the expanded conformational ensemble is fully populated, steric constraints may be the dominant constraints on the statistical conformation. Because the ensemble-averaged dimensions of both statistical conformations are so similar, it is unlikely that the distinct populations could be distinguished by means of other biophysical techniques like size-exclusion chromatography or glycerol gradient sedimentation. However, future SAXS studies of FN^{III}(1-2) at higher NaCl concentrations could validate this model.

5.5 Methods

5.5.1 SAXS sample preparation:

The proteins used in this study were expressed and purified by Riddhi Shah and Tomoo Ohashi. Following purification the proteins were flash frozen in liquid nitrogen and stored at -80˚C. The protein samples The proteins were thawed and dialyzed against 0.01 M HEPES, pH 7.4, 0.001 M DTT, and in either 0M , 0.15M or 0.50 M NaCl for 6 hours at room temperature using a 3500 Da MWCO micro-dialysis unit (Pierce).

SAXS data of the His-FN^{III}(1-2) constructs were collected at the Advanced Light Source at Argonne National Labs. For ALS data collection, samples were centrifuged at 16,000 x g for 20 minutes and then the concentration of each sample was calculated by A_{280}. Samples were diluted to a concentration of 5 mg/ml, 2.5 mg/ml or 1.25 mg/ml using dialysate. The samples were stored at 4° C for no more than 24 hours.

SAXS data of the FN^{III}(1-2) constructs were collected at Duke University on the Ganesha system. Just prior to data collection each sample was centrifuged at 16,000 x g for 20 minutes and the concentration was calculated by A_{280}. Samples were diluted to a final concentration of 1 mg/ml, using dialysate.

5.5.2 SAXS data acquisition and analysis

Data were collected at beamline 12.3.1 (SIBYLS) at the Advanced Light Source, Lawrence Berkeley National Labs, and at Duke University on the Ganesha system.

At the SIBYLS beamline, 25 μL of 5mg/ml, 2.5 mg/ml or 1.25 mg/ml protein samples were loaded into a sample cell and then exposed for 0.5, 1, or 4 seconds at an energy of 12 kEV, with a sample-to-detector distance of 1.5 M, corresponding to a q-range of 0.01 - 0.32 $Å^{-1}$. Data were collected from an identical buffer sample, using dialysate from the equilibrium dialysis, for each protein sample using identical data collection conditions. All data were collected at 10° C. Beamline specific software was used to reduce the data and subtract the buffer signal to generate final scattering data for each protein sample.

On the Ganesha system data 75 uLof 1.1mg/ml protein samples were loaded into the sample capillary and then exposed for 5.5 hours at an energy of 8.027 kEV, at a q-range of 0.015 – 0.25 $Å^{-1}$. Data from identical buffer samples from the dialysate were collected for each protein sample. All data were collected at 10° C. Data were reduced using the SAXSGui package (www.saxsgui.com). 1333 individual data-sets of 15 second exposure each for each protein were averaged in SAXSGui and the buffer signal was subtracted from the data signal using PRIMUS (Konarev et al., 2003) to generate final scattering data for each protein sample.

Guinier analysis for each construct was performed using PRIMUS to determine the radius of gyration and I(0). The barbell polymer model was fit to the scattering data using a nonlinear least squares fitting algorithm implemented in Mathematica 9. 95% confidence intervals and standard errors were calculated in Mathematica 9. Comparison of polymer models and fit of the models to the experimental SAXS data was performed in SASView. Calculation of the χ^2 statistic for all models was performed in Mathematica 9 using the reduced χ^2 statistic reported by Svergun in 1995 (Svergun et al., 1995).

5.5.3 Ensemble Modeling

Calculation of the aggregate scattering profile from the NMR structural ensemble was performed in CRYSOL (Svergun et al., 1995) and constrained against the SAXS data. Calculation of the χ^2 statistic of the fit of the aggregate scattering curve to the data was performed in Mathematica 9.

Generation of the 10,000 member parent ensemble was performed in RanCH (part of the EOM suite of programs (Bernado et al., 2007)). 2HA1.1 was used to generate the pdb model of the rigid domains. RanCH generated 10,000 structures in random coil mode. The experimental SAXS data was used to constrain the calculation of the minimal ensembles in GAJOE. 1000 generations were performed for each GAJOE cycle, and the best fit ensembles from 100 cycles were compared, and the ensemble with the lowest χ2 statistic was reported.

6. Conclusions and Future Directions

In the preceding chapters, I have highlighted the usefulness of polymer physics analysis and the judicious use of ensemble modeling to describe the statistical conformation of highly-flexible proteins. After providing an introduction to SAXS, I reviewed the common modeling methods used to interpret SAXS data and their benefits and limitations. Then I showed how a SAXS analysis of four different proteins can yield significant information about their solution conformations. In all these examples, a polymer modeling step was added to the traditional SAXS analysis protocol, and I showed how this analysis could complement existing ensemble and rigid body modeling methods. More importantly, I showed how a polymer physics model of the SAXS data could provide statistical information about the conformation of a protein that was minimally parameterized and preferable to ensemble-based modeling, which may over-fit the experimental data.

Though SAXS is a low-information-content technique, it is a tool that can address some structural problems better than standard NMR or crystallographic methods. It is a snapshot of the thermodynamic shape of an ensemble that can be applied to a wide variety of solution conditions, temperature, and particle sizes, some of which may not be accessible in x-ray crystallography or NMR. In flexible systems, the SAXS signal does not broaden or attenuate like NMR signals so it truly captures structural information about the entire thermodynamic state of the flexible protein.

As SAXS becomes more accessible to the structural biology community, it is developing into a mainstream method used to validate structural models obtained from NMR or crystallography studies. It is also used to construct *ab initio* models of proteins and protein complexes where no atomistic structure exists. However, as the research presented here highlights, describing the "structure" of a protein by small-angle scattering is not straightforward. Care must be taken to insure that the model is consistent with the low information content of the technique. Minimally parameterized models, such as polymer physics models, more accurately reflect the information content of the data than highly parameterized methods that enumerate a minimal ensemble. Since SAXS analysis is becoming more automated and "black-box" (Franke, 2012; Hura et al., 2009), it is more important than ever to develop methods that validate protein structures resulting from a SAXS analysis. It is also important when a SAXS-based model is published, that the inherent limitations of the model are understood and discussed. I have attempted to do this in this manuscript.

In the future, the integration of SAXS analysis within NMR and crystallography structure refinement programs will allow researchers to take advantage of the unique information that comes from SAXS data while at the same time using these other techniques to generate and validate an atomistic model representative of the solution conformation of the molecule. SAXS constraints are already implemented in two NMR refinement programs: XPLOR-NIH (Schwieters et al., 2003) and Flexible-Meccano

(Ozenne et al., 2012). However, the only SAXS-derived information that these two programs use in structure refinement is the ensemble-averaged radius of gyration. As I have shown, much more information can be determined about the statistical conformation than solely the average R_g. I would like to explore the possibility of using other parameters determined from a polymer physics SAXS analysis in NMR structure refinement of highly flexible proteins. For example, if a protein can be described as an excluded volume Gaussian coil, then the Flory coefficient and persistence length can be two more parameters that are used in structural refinement.

Another area of future SAXS development is in the prediction of scattering profiles from atomistic models. As I showed in Chapter 2, current prediction programs cannot predict the scattering profile of an atomistic model when the program does not use the experimental SAXS data as a constraint. Instead, the software uses an adjustable hydration layer parameter to optimize the fit of the theoretical profile to the data. In theory, however, it should be possible to generate theoretical scattering profile unconstrained by the data that accurately predicts the experimental data. This is one area of research I hope to explore during my postdoctoral training.

In order for the prediction methods to be improved, researchers need access to a large amount of experimental data on which to test their algorithms. Unfortunately, most of the prediction programs are only tested on *in silico* simulated SAXS results or on a small number (<10) of experimental datasets. Since 2000 there has been an explosion in the number of publications per year that contain a scattering analysis (Figure 40). However, there are only 77 SAXS datasets of proteins and nucleic acids available in public databases (www.bioisis.net, (Varadi et al., 2013). Until deposition of data in public databases is mandated, it will be very hard to test any type of SAXS structural refinement software against experimental data.

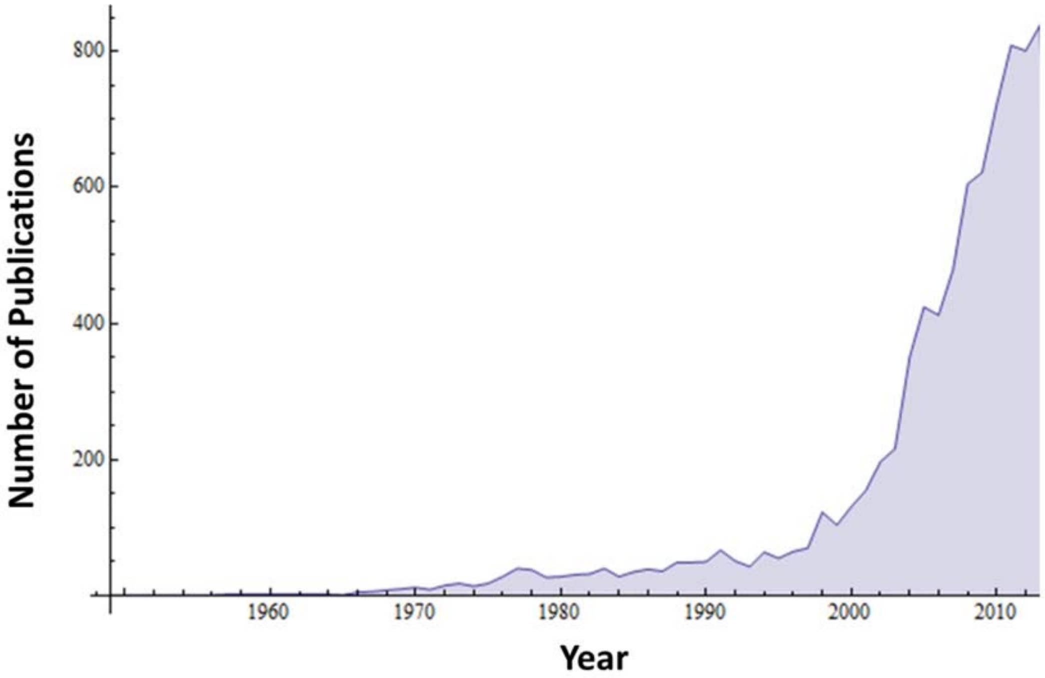

Figure 40: The number of publications per year that contain the key words "small angle" and "scattering" in the PubMed database (www.ncbi.nlm.nih.gov/pubmed/).

In conclusion, SAXS analysis of highly flexible proteins can yield important information about the statistical conformation that is inaccessible by other structural biology techniques. However, when fitting models to the experimental SAXS data, it is important to start with minimally parameterized models and only increase the number of parameters in the model when doing so is supported by the data.

Appendix A: Transformation of Cartesian coordinates to spherical polar coordinates

The purpose of this section is to derive the transformation of Cartesian coordinates to spherical polar coordinates in order to prove equation 1.21.

The relationship between the coordinate systems is:

$$x = r\sin\theta\cos\Phi \qquad \text{A.1}$$

$$y = r\sin\theta\sin\Phi \qquad \text{A.2}$$

$$z = r\cos\theta \qquad \text{A.3}$$

Now we define x,y,z as functions of r,θ,φ:

$$x = f(\mathrm{r}, \theta, \Phi) = r\sin\theta\cos\Phi \qquad \text{A.4}$$

$$y = g(\mathrm{r}, \theta, \Phi) = r\sin\theta\sin\Phi \qquad \text{A.5}$$

$$z = h(\mathrm{r}, \theta, \Phi) = r\cos\theta \qquad \text{A.6}$$

The volume element, dV in Cartesian coordinates is

$$dV = dx\,dy\,dz \qquad \text{A.7}$$

And the volume is calculated using a triple integral:

$$Volume = \int\int\int F(x, y, z)dx\,dy\,dz \qquad \text{A.8}$$

Similarly, the volume element in spherical polar coordinates is calculated using a triple integral:

$$Volume = \int\int\int G(\mathrm{r}, \theta, \Phi)|J|dr\,d\theta\,d\Phi \qquad \text{A.9}$$

In equation 9, dV therefore is:

$$dV = |J| dr\ d\theta\ d\Phi \tag{A.10}$$

$|J|$ is the absolute value of the Jacobian that transforms from the Cartesian coordinate to the spherical polar coordinate system. In order to solve for dV, it is necessary to derive $|J|$. First we state the determinant of the matrix that transforms $dx\ dy\ dz$ into $dr\ d\theta\ d\Phi$:

$$dx = \frac{\delta f(\mathrm{r},\theta,\Phi)}{\delta r} dr + \frac{\delta f(\mathrm{r},\theta,\Phi)}{\delta \theta} d\theta + \frac{\delta f(\mathrm{r},\theta,\Phi)}{\delta \Phi} d\Phi \tag{A.11}$$

$$dy = \frac{\delta g(\mathrm{r},\theta,\Phi)}{\delta r} dr + \frac{\delta g(\mathrm{r},\theta,\Phi)}{\delta \theta} d\theta + \frac{\delta g(\mathrm{r},\theta,\Phi)}{\delta \Phi} d\Phi \tag{A.12}$$

$$dz = \frac{\delta h(\mathrm{r},\theta,\Phi)}{\delta r} dr + \frac{\delta h(\mathrm{r},\theta,\Phi)}{\delta \theta} d\theta + \frac{\delta h(\mathrm{r},\theta,\Phi)}{\delta \Phi} d\Phi \tag{A.13}$$

The Jacobian matrix is set up as:

$$|J| = det \begin{pmatrix} \frac{\delta f}{\delta r} & \frac{\delta f}{\delta \theta} & \frac{\delta f}{\delta \Phi} \\ \frac{\delta g}{\delta r} & \frac{\delta g}{\delta \theta} & \frac{\delta g}{\delta \Phi} \\ \frac{\delta h}{\delta r} & \frac{\delta h}{\delta \theta} & \frac{\delta h}{\delta \Phi} \end{pmatrix} \tag{A.14}$$

To calculate the determinate:

$$det \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = (aei + bfg + cdh) - (ceg + bdi + afh)$$

So,

$$det \begin{vmatrix} \frac{\delta f}{\delta r} & \frac{\delta f}{\delta \theta} & \frac{\delta f}{\delta \Phi} \\ \frac{\delta g}{\delta r} & \frac{\delta g}{\delta \theta} & \frac{\delta g}{\delta \Phi} \\ \frac{\delta h}{\delta r} & \frac{\delta h}{\delta \theta} & \frac{\delta h}{\delta \Phi} \end{vmatrix} = \left(\frac{\delta f}{\delta r}\frac{\delta g}{\delta \theta}\frac{\delta h}{\delta \Phi} + \frac{\delta f}{\delta \theta}\frac{\delta g}{\delta \Phi}\frac{\delta h}{\delta r} + \frac{\delta f}{\delta \Phi}\frac{\delta g}{\delta r}\frac{\delta h}{\delta \theta}\right) - \left(\frac{\delta f}{\delta \Phi}\frac{\delta g}{\delta \theta}\frac{\delta h}{\delta r} + \frac{\delta f}{\delta \theta}\frac{\delta g}{\delta r}\frac{\delta h}{\delta \Phi} + \frac{\delta f}{\delta r}\frac{\delta g}{\delta \Phi}\frac{\delta h}{\delta \theta}\right) \tag{A.15}$$

The following derivatives are used:

$$\left(\frac{\delta f}{\delta r}\right)_{\theta,\varPhi} = \sin\theta\cos\varPhi \qquad \left(\frac{\delta f}{\delta \theta}\right)_{r,\varPhi} = r\cos\theta\cos\varPhi \qquad \left(\frac{\delta f}{\delta \varPhi}\right)_{r,\theta} = -r\sin\theta\sin\varPhi$$

$$\left(\frac{\delta g}{\delta r}\right)_{\theta,\varPhi} = \sin\theta\sin\varPhi \qquad \left(\frac{\delta g}{\delta \theta}\right)_{r,\varPhi} = r\cos\theta\sin\varPhi \qquad \left(\frac{\delta g}{\delta \varPhi}\right)_{r,\theta} = r\sin\theta\cos\varPhi$$

$$\left(\frac{\delta h}{\delta r}\right)_{\theta,\varPhi} = \cos\theta \qquad \left(\frac{\delta h}{\delta \theta}\right)_{r,\varPhi} = -r\sin\theta \qquad \left(\frac{\delta h}{\delta \varPhi}\right)_{r,\theta} = 0$$

Thus,

$$\begin{aligned} |J| = {} & (\sin\theta\cos\varPhi * r\cos\theta\sin\varPhi * 0 + r\cos\theta\cos\varPhi * r\sin\theta\cos\varPhi * \cos\theta + -r\sin\theta\sin\varPhi \\ & * \sin\theta\sin\varPhi * -r\sin\theta) - (-r\sin\theta\sin\varPhi * r\cos\theta\sin\varPhi * \cos\theta \\ & + r\cos\theta\cos\varPhi * \sin\theta\sin\varPhi * 0 + \sin\theta\cos\varPhi * r\sin\theta\cos\varPhi - r\sin\theta) \end{aligned}$$

$$\begin{aligned} |J| = {} & (0 + r^2\sin\theta\cos^2\theta\cos^2\varPhi + r^2\sin^3\theta\sin^2\varPhi) - (-r^2\sin\theta\cos^2\theta\sin^2\varPhi + 0 \\ & - r^2\sin^3\theta\cos^2\varPhi) \end{aligned}$$

$$|J| = r^2\sin\theta\,(\cos^2\theta\cos^2\varPhi + \sin^2\theta\sin^2\varPhi + \cos^2\theta\sin^2\varPhi + \sin^2\theta\cos^2\varPhi) \tag{A.16}$$

Using the identity $\sin^2\theta + \cos^2\theta = 1$:

$$|J| = r^2\sin\theta\,(\cos^2\theta\cos^2\varPhi + \sin^2\varPhi\,(\sin^2\theta + \cos^2\theta) + \sin^2\theta\cos^2\varPhi) \tag{A.17}$$

$$|J| = r^2\sin\theta\,(\cos^2\theta\cos^2\varPhi + \sin^2\theta\cos^2\varPhi + \sin^2\varPhi) \tag{A.18}$$

$$|J| = r^2\sin\theta\,(\cos^2\varPhi\,(\cos^2\theta + \sin^2\theta) + \sin^2\varPhi) \tag{A.19}$$

$$|J| = r^2 \sin\theta \,(\cos^2 \Phi + \sin^2 \Phi) \qquad \text{A.20}$$

$$|J| = r^2 \sin\theta \qquad \text{A.21}$$

In equation 1.21, we are integrating over the whole volume of **r**, so,

$$d\boldsymbol{r} = dV = dx\; dy\; dz$$

When this is transformed into spherical polar coordinates,

$$d\boldsymbol{r} = dV = r^2 \sin\theta \; dr\; d\theta\; d\Phi \qquad \text{A.22}$$

References

Baldwin, R.L. (2002). A new perspective on unfolded proteins. Advances in protein chemistry *62*, 361-367.

Berlin, K., Castaneda, C.A., Schneidman-Duhovny, D., Sali, A., Nava-Tudela, A., and Fushman, D. (2013). Recovering a representative conformational ensemble from underdetermined macromolecular structural data. Journal of the American Chemical Society *135*, 16595-16609.

Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M., and Svergun, D.I. (2007). Structural characterization of flexible proteins using small-angle X-ray scattering. Journal of the American Chemical Society *129*, 5656-5664.

Bernado, P., and Svergun, D.I. (2012). Analysis of intrinsically disordered proteins by small-angle X-ray scattering. Methods in molecular biology *896*, 107-122.

Bevington, P.R., and Robinson, D.K. (2003). Data reduction and error analysis for the physical sciences, 3rd edn (Boston: McGraw-Hill).

Biersmith, B.H., Hammel, M., Geisbrecht, E.R., and Bouyain, S. (2011). The immunoglobulin-like domains 1 and 2 of the protein tyrosine phosphatase LAR adopt an unusual horseshoe-like conformation. Journal of molecular biology *408*, 616-627.

Bracken, C. (2001). NMR spin relaxation methods for characterization of disorder and folding in proteins. Journal of molecular graphics & modelling *19*, 3-12.

Burchard, W., and Kajiwara, K. (1970). The Statistics of Stiff Chain Molecules. I. The Particle Scattering Factor. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences *316*, 185-199.

Chi, Q., Wang, G., and Jiang, J. (2013). The persistence length and length per base of single-stranded DNA obtained from fluorescence correlation spectroscopy measurements using mean field theory. Physica A: Statistical Mechanics and its Applications *392*, 1072-1079.

Curtis, J.E., Raghunandan, S., Nanda, H., and Krueger, S. (2012). SASSIE: A program to study intrinsically disordered biological molecules and macromolecular ensembles using experimental scattering restraints. Computer Physics Communications *183*, 382-389.

D. Walther, F.E.C., S. Doniach (2000). Reconstruction of low resolution three-dimensional density maps from one-dimensional small-angle x-ray solution scattering data for biomolecules. J Appl Cryst *33*, 350-363.

Das, R.K., and Pappu, R.V. (2013). Conformations of intrinsically disordered proteins are influenced by linear sequence distributions of oppositely charged residues. Proceedings of the National Academy of Sciences of the United States of America *110*, 13392-13397.

De Biasio, A., Ibanez de Opakua, A., Cordeiro, T.N., Villate, M., Merino, N., Sibille, N., Lelli, M., Diercks, T., Bernado, P., and Blanco, F.J. (2014). p15(PAF) Is an Intrinsically Disordered Protein with Nonrandom Structural Preferences at Sites of Interaction with Other Proteins. Biophysical journal *106*, 865-874.

Deisenhofer, J. (1981). Crystallographic refinement and atomic models of a human Fc fragment and its complex with fragment B of protein A from Staphylococcus aureus at 2.9- and 2.8-A resolution. Biochemistry *20*, 2361-2370.

Dobrynin, A.V., Rubinstein, M., and Obukhov, S.P. (1996). Cascade of transitions of polyelectrolytes in poor solvents. Macromolecules *29*, 2974-2979.

Drenth, J., and Mesters, J. (2007). Principles of protein X-ray crystallography, 3rd edn (New York: Springer).

Durand, D., Vives, C., Cannella, D., Perez, J., Pebay-Peyroula, E., Vachette, P., and Fieschi, F. (2010). NADPH oxidase activator p67(phox) behaves in solution as a multidomain protein with semi-flexible linkers. Journal of structural biology *169*, 45-53.

Engelman, D.M., and Moore, P.B. (1972). A new method for the determination of biological quarternary structure by neutron scattering. Proceedings of the National Academy of Sciences of the United States of America *69*, 1997-1999.

Flory, P.J. (1953). Principles of polymer chemistry (Ithaca,: Cornell University Press).

Fontana, A., de Laureto, P.P., Spolaore, B., Frare, E., Picotti, P., and Zambonin, M. (2004). Probing protein structure by limited proteolysis. Acta biochimica Polonica *51*, 299-321.

Franke, A., Kikhney, A.G., Svergun, D. (2012). Automated acquisition and analysis of small-angle x-ray scattering data. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment *689*, 52-59.

Fujita, H. (1990). Polymer solutions (Amsterdam ;New York, NY, U.S.A.: Elsevier ; Distributors for the U.S. and Canada, Elsevier Science Pub.).

Glatter, O., and Kratky, O. (1982). Small angle x-ray scattering (London ; New York: Academic Press).

Gomez, M.I., Lee, A., Reddy, B., Muir, A., Soong, G., Pitt, A., Cheung, A., and Prince, A. (2004). Staphylococcus aureus protein A induces airway epithelial inflammatory responses by activating TNFR1. Nature medicine *10*, 842-848.

Graille, M., Stura, E.A., Corper, A.L., Sutton, B.J., Taussig, M.J., Charbonnier, J.B., and Silverman, G.J. (2000). Crystal structure of a Staphylococcus aureus protein A domain complexed with the Fab fragment of a human IgM antibody: structural basis for recognition of B-cell receptors and superantigen activity. Proceedings of the National Academy of Sciences of the United States of America *97*, 5399-5404.

Grishaev, A., Guo, L., Irving, T., and Bax, A. (2010). Improved fitting of solution X-ray scattering data to macromolecular structures and structural ensembles by explicit water modeling. Journal of the American Chemical Society *132*, 15484-15486.

Guinier, A., and Fournet, G.r. (1955). Small-angle scattering of X-rays (New York,: Wiley).

Gumerov, N.A., Berlin, K., Fushman, D., and Duraiswami, R. (2012). A hierarchical algorithm for fast Debye summation with applications to small angle scattering. Journal of computational chemistry *33*, 1981-1996.

Guss, B., Uhlen, M., Nilsson, B., Lindberg, M., Sjoquist, J., and Sjodahl, J. (1984). Region X, the cell-wall-attachment part of staphylococcal protein A. European journal of biochemistry / FEBS *138*, 413-420.

Hagerman, P.J. (1988). Flexibility of DNA. Annual review of biophysics and biophysical chemistry *17*, 265-286.

Hammel, M. (2012). Validation of macromolecular flexibility in solution by small-angle X-ray scattering (SAXS). European biophysics journal : EBJ *41*, 789-799.

Hammouda, B. (1993). SANS from homogenous polymer mixtures: A unified overview. Advances in Polymer Science *106*, 87-133.

Hartleib, J., Kohler, N., Dickinson, R.B., Chhatwal, G.S., Sixma, J.J., Hartford, O.M., Foster, T.J., Peters, G., Kehrel, B.E., and Herrmann, M. (2000). Protein A is the von Willebrand factor binding protein on Staphylococcus aureus. Blood *96*, 2149-2156.

Hernandez, G., and LeMaster, D.M. (2009). NMR analysis of native-state protein conformational flexibility by hydrogen exchange. Methods in molecular biology *490*, 285-310.

Hjelm, R.P. (1985). The Small-Angle Approximation of X-ray and Neutron Scatter from Rigid Rods of Non-Uniform Cross-Section and Finite Length. Journal of Applied Crystallography *18*, 452-460.

Hura, G.L., Budworth, H., Dyer, K.N., Rambo, R.P., Hammel, M., McMurray, C.T., and Tainer, J.A. (2013). Comprehensive macromolecular conformations mapped by quantitative SAXS analyses. Nature methods *10*, 453-454.

Hura, G.L., Menon, A.L., Hammel, M., Rambo, R.P., Poole, F.L., 2nd, Tsutakawa, S.E., Jenney, F.E., Jr., Classen, S., Frankel, K.A., Hopkins, R.C., *et al.* (2009). Robust, high-throughput solution structural analyses by small angle X-ray scattering (SAXS). Nature methods *6*, 606-612.

Jacques, D.A., Guss, J.M., Svergun, D.I., and Trewhella, J. (2012). Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution. Acta crystallographica Section D, Biological crystallography *68*, 620-626.

Kamide, K., and Dobashi, T. (2000). Physical chemistry of polymer solutions : theoretical background, 1st edn (Amsterdam ; New York: Elsevier Science BV).

Karuri, N.W., Lin, Z., Rye, H.S., and Schwarzbauer, J.E. (2009). Probing the conformation of the fibronectin III1-2 domain by fluorescence resonance energy transfer. The Journal of biological chemistry *284*, 3445-3452.

Kaya, H. (2004). Scattering from cylinders with globular end-caps. Journal of Applied Crystallography *37*, 223-230.

Kline, S.R. (2006). Reduction and analysis of SANS and USANS data using Igor Pro. Journal of Applied Crystallography *39*, 895-900.

Koch, M.H., Vachette, P., and Svergun, D.I. (2003). Small-angle scattering: a view on the properties, structures and structural changes of biological macromolecules in solution. Quarterly reviews of biophysics *36*, 147-227.

Kohn, J.E., Millett, I.S., Jacob, J., Zagrovic, B., Dillon, T.M., Cingel, N., Dothager, R.S., Seifert, S., Thiyagarajan, P., Sosnick, T.R., *et al.* (2004). Random-coil behavior and the dimensions of chemically unfolded proteins. Proceedings of the National Academy of Sciences of the United States of America *101*, 12491-12496.

Konarev, P.V., Volkov, V.V., Sokolova, A.V., Koch, M.H.J., and Svergun, D.I. (2003). PRIMUS: a Windows PC-based system for small-angle scattering data analysis. Journal of Applied Crystallography *36*, 1277-1282.

Konig, S., Svergun, D., Koch, M.H., Hubner, G., and Schellenberger, A. (1993). The influence of the effectors of yeast pyruvate decarboxylase (PDC) on the conformation of the dimers and tetramers and their pH-dependent equilibrium. European biophysics journal : EBJ *22*, 185-194.

Kornblihtt, A.R., Umezawa, K., Vibe-Pedersen, K., and Baralle, F.E. (1985). Primary structure of human fibronectin: differential splicing may generate at least 10 polypeptides from a single gene. The EMBO journal *4*, 1755-1759.

Krzeminski, M., Marsh, J.A., Neale, C., Choy, W.Y., and Forman-Kay, J.D. (2013). Characterization of disordered proteins with ENSEMBLE. Bioinformatics *29*, 398-399.

Kurata, M., Yamakawa, H., and Teramoto, E. (1958). Theory of Dilute Polymer Solution. I. Excluded Volume Effect. The Journal of chemical physics *28*, 785.

Lairez, D., Pauthe, E., and Pelta, J. (2003). Refolding of a high molecular weight protein: salt effect on collapse. Biophysical journal *84*, 3904-3916.

Leahy, D.J., Aukhil, I., and Erickson, H.P. (1996). 2.0 A crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. Cell *84*, 155-164.

Lemmon, C.A., Ohashi, T., and Erickson, H.P. (2011). Probing the folded state of fibronectin type III domains in stretched fibrils by measuring buried cysteine accessibility. The Journal of biological chemistry *286*, 26375-26382.

Levinthal, C. (1968). Are there pathways for protein folding. The Journal of chemical physics *65*, 44-45.

Liu, H., Morris, R.J., Hexemer, A., Grandison, S., and Zwart, P.H. (2012). Computation of small-angle scattering profiles with three-dimensional Zernike polynomials. Acta crystallographica Section A, Foundations of crystallography *68*, 278-285.

Lofdahl, S., Guss, B., Uhlen, M., Philipson, L., and Lindberg, M. (1983). Gene for staphylococcal protein A. Proceedings of the National Academy of Sciences of the United States of America *80*, 697-701.

Mackerell, A.D., Jr., Feig, M., and Brooks, C.L., 3rd (2004). Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. Journal of computational chemistry *25*, 1400-1415.

Merino, N., Toledo-Arana, A., Vergara-Irigaray, M., Valle, J., Solano, C., Calvo, E., Lopez, J.A., Foster, T.J., Penades, J.R., and Lasa, I. (2009). Protein A-mediated multicellular behavior in Staphylococcus aureus. Journal of bacteriology *191*, 832-843.

Mertens, H.D., and Svergun, D.I. (2010). Structural characterization of proteins and complexes using small-angle X-ray solution scattering. Journal of structural biology *172*, 128-141.

Moks, T., Abrahmsen, L., Nilsson, B., Hellman, U., Sjoquist, J., and Uhlen, M. (1986). Staphylococcal protein A consists of five IgG-binding domains. European journal of biochemistry / FEBS *156*, 637-643.

Moore, P.B., Capel, M., Kjeldgaard, M., and Engelman, D.M. (1986). Quaternary Organization of the 30S Ribosomal Subunit of Escherichia Coli. Biophysical journal *49*, 13-15.

Nguyen, T., Ghebrehiwet, B., and Peerschke, E.I. (2000). Staphylococcus aureus protein A recognizes platelet gC1qR/p33: a novel mechanism for staphylococcal interactions with platelets. Infection and immunity *68*, 2061-2068.

Ohashi, T., and Erickson, H.P. (2011). Fibronectin aggregation and assembly: the unfolding of the second fibronectin type III domain. The Journal of biological chemistry *286*, 39188-39199.

Ozenne, V., Bauer, F., Salmon, L., Huang, J.R., Jensen, M.R., Segard, S., Bernado, P., Charavay, C., and Blackledge, M. (2012). Flexible-meccano: a tool for the generation of explicit ensemble descriptions of intrinsically disordered proteins and their associated experimental observables. Bioinformatics *28*, 1463-1470.

Palmqvist, N., Foster, T., Tarkowski, A., and Josefsson, E. (2002). Protein A is a virulence factor in Staphylococcus aureus arthritis and septic death. Microbial pathogenesis *33*, 239-249.

Pedersen, J.S. (1996). Scattering functions of semiflexible polymers with and without excluded volume effects. Macromolecules *29*, 7602-7612.

Pedersen, J.S., Laso, M., and Schurtenberger, P. (1996). Monte Carlo study of excluded volume effects in wormlike micelles and semiflexible polymers. Physical review E, Statistical physics, plasmas, fluids, and related interdisciplinary topics *54*, R5917-R5920.

Pelikan, M., Hura, G.L., and Hammel, M. (2009). Structure and flexibility within proteins as identified through small angle X-ray scattering. General physiology and biophysics *28*, 174-189.

Pelta, J., Berry, H., Fadda, G.C., Pauthe, E., and Lairez, D. (2000). Statistical conformation of human plasma fibronectin. Biochemistry *39*, 5146-5154.

Petersen, T.E., Thogersen, H.C., Skorstengaard, K., Vibe-Pedersen, K., Sahl, P., Sottrup-Jensen, L., and Magnusson, S. (1983). Partial primary structure of bovine plasma fibronectin: three types of internal homology. Proceedings of the National Academy of Sciences of the United States of America *80*, 137-141.

Petoukhov, M.V., and Svergun, D.I. (2005). Global rigid body modeling of macromolecular complexes against small-angle scattering data. Biophysical journal *89*, 1237-1250.

Potts, J.R., and Campbell, I.D. (1994). Fibronectin structure and assembly. Current opinion in cell biology *6*, 648-655.

Putnam, C.D., Hammel, M., Hura, G.L., and Tainer, J.A. (2007). X-ray solution scattering (SAXS) combined with crystallography and computation: defining accurate macromolecular structures, conformations and assemblies in solution. Quarterly reviews of biophysics *40*, 191-285.

Rambo, R.P., and Tainer, J.A. (2011). Characterizing flexible and intrinsically unstructured biological macromolecules by SAS using the Porod-Debye law. Biopolymers *95*, 559-571.

Rambo, R.P., and Tainer, J.A. (2013). Accurate assessment of mass, models and resolution by small-angle scattering. Nature *496*, 477-481.

Ravikumar, K.M., Huang, W., and Yang, S. (2013). Fast-SAXS-pro: a unified approach to computing SAXS profiles of DNA, RNA, protein, and their complexes. The Journal of chemical physics *138*, 024112.

Receveur-Brechot, V., and Durand, D. (2012). How random are intrinsically disordered proteins? A small angle scattering perspective. Current protein & peptide science *13*, 55-75.

Roe, R.J. (2000). Methods of X-ray and neutron scattering in polymer science (New York: Oxford University Press).

Rozycki, B., Kim, Y.C., and Hummer, G. (2011). SAXS ensemble refinement of ESCRT-III CHMP3 conformational transitions. Structure *19*, 109-116.

Rubinstein, M., and Colby, R.H. (2003). Polymer physics (Oxford ; New York: Oxford University Press).

Schneewind, O., Fowler, A., and Faull, K.F. (1995). Structure of the cell wall anchor of surface proteins in Staphylococcus aureus. Science *268*, 103-106.

Schneidman-Duhovny, D., Hammel, M., and Sali, A. (2010). FoXS: a web server for rapid computation and fitting of SAXS profiles. Nucleic acids research *38*, W540-544.

Schwarzbauer, J.E. (1991). Identification of the fibronectin sequences required for assembly of a fibrillar matrix. The Journal of cell biology *113*, 1463-1473.

Schweins, R., and Huber, K. (2004). Particle scattering factor of pearl necklace chains. Macromolecular Symposia *211*, 25-42.

Schwieters, C.D., Kuszewski, J.J., Tjandra, N., and Clore, G.M. (2003). The Xplor-NIH NMR molecular structure determination package. Journal of magnetic resonance *160*, 65-73.

Sechler, J.L., Rao, H., Cumiskey, A.M., Vega-Colon, I., Smith, M.S., Murata, T., and Schwarzbauer, J.E. (2001). A novel fibronectin binding site required for fibronectin fibril growth during matrix assembly. The Journal of cell biology *154*, 1081-1088.

Serdyuk, I.N., Tsalkova, T.N., Svergun, D.I., and Izotova, T.D. (1987). Determination of radii of gyration of particles by small-angle neutron scattering: calculation of the effect of aggregates. Journal of molecular biology *194*, 126-128.

Singh, P., Carraher, C., and Schwarzbauer, J.E. (2010). Assembly of fibronectin extracellular matrix. Annual review of cell and developmental biology *26*, 397-419.

Sjoholm, I., Ekenas, A.K., and Sjoquist, J. (1972). Protein A from Staphylococcus aureus. Acetylation of protein A with acetylimidazole. European journal of biochemistry / FEBS *29*, 455-460.

Starovasnik, M.A., Skelton, N.J., O'Connell, M.P., Kelley, R.F., Reilly, D., and Fairbrother, W.J. (1996). Solution structure of the E-domain of staphylococcal protein A. Biochemistry *35*, 15558-15569.

Stollar, E.J., Lin, H., Davidson, A.R., and Forman-Kay, J.D. (2012). Differential dynamic engagement within 24 SH3 domain: peptide complexes revealed by co-linear chemical shift perturbation analysis. PloS one *7*, e51282.

Svergun, D., Barberato, C., and Koch, M.H.J. (1995). CRYSOL - A program to evaluate x-ray solution scattering of biological macromolecules from atomic coordinates. Journal of Applied Crystallography *28*, 768-773.

Svergun, D.I. (1992). Determination of the Regularization Parameter in Indirect-Transform Methods Using Perceptual Criteria. Journal of Applied Crystallography *25*, 495-503.

Svergun, D.I., Feĭgin, L.A., and Taylor, G.W. (1987). Structure analysis by small-angle x-ray and neutron scattering (New York: Plenum Press).

Tjioe, E.a.W.T.H. (2007). ORNL_SAS: software for calculation of small-angle scattering intensities of proteins and protein complexes. J Appl Cryst *40*, 782-785.

Tokuriki, N., and Tawfik, D.S. (2009). Protein dynamism and evolvability. Science *324*, 203-207.

Vachette, P., Koch, M.H., and Svergun, D.I. (2003). Looking behind the beamstop: X-ray solution scattering studies of structure and conformational changes of biological macromolecules. Methods in enzymology *374*, 584-615.

Vakonakis, I., Staunton, D., Rooney, L.M., and Campbell, I.D. (2007). Interdomain association in fibronectin: insight into cryptic sites and fibrillogenesis. The EMBO journal *26*, 2575-2583.

Varadi, M., Kosol, S., Lebrun, P., Valentini, E., Blackledge, M., Dunker, A.K., Felli, I.C., Forman-Kay, J.D., Kriwacki, R.W., Pierattelli, R., *et al.* (2013). pE-DB: a database of structural ensembles of intrinsically disordered and of unfolded proteins. Nucleic acids research.

Virtanen, J.J., Makowski, L., Sosnick, T.R., and Freed, K.F. (2010). Modeling the hydration layer around proteins: HyPred. Biophysical journal *99*, 1611-1619.

Wignall, G.D., Ballard, D.G.H., and Schelten, J. (1974). Measurements of Persistence Length and Temperature-Dependence of Radius of Gyration in Bulk Atactic Polystyrene. Eur Polym J *10*, 861-865.

Williams, C.K., Vaithiyalingam, S., Hammel, M., Pipas, J., and Chazin, W.J. (2012). Binding to retinoblastoma pocket domain does not alter the inter-domain flexibility of the J domain of SV40 large T antigen. Archives of biochemistry and biophysics *518*, 111-118.

Yang, S., Blachowicz, L., Makowski, L., and Roux, B. (2010). Multidomain assembled states of Hck tyrosine kinase in solution. Proceedings of the National Academy of Sciences of the United States of America *107*, 15757-15762.

Yang, S., Park, S., Makowski, L., and Roux, B. (2009). A rapid coarse residue-based computational method for x-ray solution scattering characterization of protein folds and multiple conformational states of large protein complexes. Biophysical journal *96*, 4449-4463.

Zheng, D., Aramini, J.M., and Montelione, G.T. (2004). Validation of helical tilt angles in the solution NMR structure of the Z domain of Staphylococcal protein A by combined analysis of residual dipolar coupling and NOE data. Protein science : a publication of the Protein Society *13*, 549-554.

Biography

Jo Anna Wiersma Capp was born in Okeechobee, Florida. She received her Associate in Arts degree from Indian River Community College and her Bachelor of Science in Biotechnology in 2006 from Florida Gulf Coast University. Upon entering graduate school in the Department of Biochemistry at Duke University, she was awarded the National Science Foundation's Graduate Research Fellowship. Jo Anna joined the Oas lab in 2012, where she studied all things SAXS.